SELF-EVOLVING SPATIAL REASONING IN VISION LANGUAGE MODELS VIA GEOMETRIC LOGIC CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have made striking progress, yet their spatial reasoning remains fragile. Models that answer an original input correctly can still fail under valid transformations with predictable answer mappings, revealing a gap between instance-level correctness and robust spatial reasoning. To address this, we propose **S**patial **A**lignment via **G**eometric **E**volution (SAGE), a self-evolving framework that improves robust spatial reasoning through geometric and linguistic duality operations. SAGE incorporates duality consistency into GRPO training, encouraging models to produce coherent answers across original and transformed inputs. SAGE co-evolves duality generation and solution, allowing the model to continually expose and address its own reasoning weaknesses. A dynamic operation pool identifies challenging operations and retires mastered ones, keeping training focused on informative duality signals. SAGE is model-agnostic, data-efficient compared to prior post-training methods, and can be applied as a lightweight adaptation stage to any existing VLM. Experiments on video and spatial reasoning benchmarks demonstrate consistent improvements over strong baselines and enhanced generalization to unseen data.

1 INTRODUCTION

Vision-Language Models (VLMs) Singh et al. (2025); Bai et al. (2025a) have advanced rapidly in recent years, evolving from basic perception toward comprehensive world understanding and multimodal reasoning. Despite this progress, their representation of spatial relationships and physical scenes remains limited Cheng et al. (2024a). When asked about object positions, motion directions, relative distances, or depth orderings, VLMs often produce confident but incorrect answers, exhibiting spatial hallucinations that undermine their reliability in embodied navigation Zheng et al. (2024) and autonomous driving Zhou et al. (2026b). Since these applications require models to reason consistently about how objects are arranged and how scenes change under viewpoint or geometric transformations, improving spatial reasoning has become a central challenge for multimodal intelligence.

Recent methods have applied Reinforcement Learning (RL) to improve VLM reasoning, including spatial and temporal reasoning tasks Zhao et al. (2025b); Pan & Liu (2026); Li et al. (2026b). These methods design reward signals to encourage desirable behaviors such as correct answers, improved temporal modeling Feng et al. (2025), or more precise geometric reasoning Wu et al. (2025b). However, existing objectives do not directly enforce consistency across paired formulations of the same spatial relation Li et al. (2026c); Zhang et al. (2026b). For example, after a horizontal flip, the correct answer to a left/right question should change predictably; if a model answers the original input correctly but fails on its flipped counterpart, it exposes a gap between answer correctness and robust spatial reasoning Kim et al. (2026). As shown in Figure 1, such transformed questions are often intuitive to humans but remain challenging for VLMs. Moreover, although related forms of transformations have been considered in prior work Fang et al. (2025); Zhou et al. (2026a), they are typically treated as predefined augmentations or task-specific perturbations rather than formalized as duality operations that can adapt to the model's evolving weaknesses.

To address this challenge, we propose **S**patial **A**lignment via **G**eometric **E**volution (SAGE), a self-evolving post-training framework that improves VLM spatial reasoning by enforcing duality

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Analysis of spatial consistency in VLMs under geometric transformations.

consistency. SAGE formalizes paired transformations as *duality operations*, including geometric changes to the visual input, such as spatial reflections, and structural changes to the query, such as option permutation. Each duality operation specifies how the input is transformed and how the corresponding answer should change, allowing the model to be trained toward answer-level equivariance. Given an original sample and its dual counterpart, SAGE encourages the model to produce logically consistent answers under the known transformation mapping. We integrate this consistency signal as an auxiliary reward within GRPO training, providing supervision beyond standard answer accuracy. Crucially, SAGE does not rely on a fixed set of transformations throughout training. Instead, it maintains a dynamic pool of duality operations that periodically probes the model for inconsistencies, identifies transformations associated with current reasoning failures, and promotes useful operations for active training. This self-evolving process allows the training distribution to adapt to the model's observed inconsistency patterns rather than relying on a manually fixed set of transformations. SAGE further uses a shared VLM for both proposing candidate duality operations and solving the resulting original–dual pairs, with deterministic validation used to verify the generated operations. This shared-parameter design avoids introducing additional trainable role-specific models while preserving the distinction between operation generation and solution. We evaluate SAGE across a broad set of video understanding and spatial reasoning benchmarks to determine whether duality consistency improves generalization beyond the specific transformations used during training. Experiments on six video understanding and seven spatial reasoning benchmarks show consistent improvements on the majority of benchmarks. These results suggest that duality consistency provides a complementary post-training signal to task-specific rewards and helps expose reasoning failures that standard answer accuracy alone may overlook.

Our technical contributions are summarized as follows:

1. We introduce *duality consistency* as a general training signal for VLM spatial reasoning. By formulating visual and query transformations together with their corresponding answer mappings, SAGE trains models toward answer-level equivariance and exposes inconsistencies that standard answer accuracy may overlook.

2. We propose SAGE, a self-evolving post-training framework that integrates duality consistency into GRPO. SAGE uses a shared model to generate and solve duality operations, while maintaining a dynamic operation pool that adaptively prioritizes transformations associated with the model's current reasoning vulnerabilities.

3. Across six video understanding benchmarks and seven spatial reasoning benchmarks, SAGE improves duality consistency and achieves competitive or superior performance using substantially less training data than prior GRPO-based methods, showing that consistency-based rewards complement task-specific post-training.

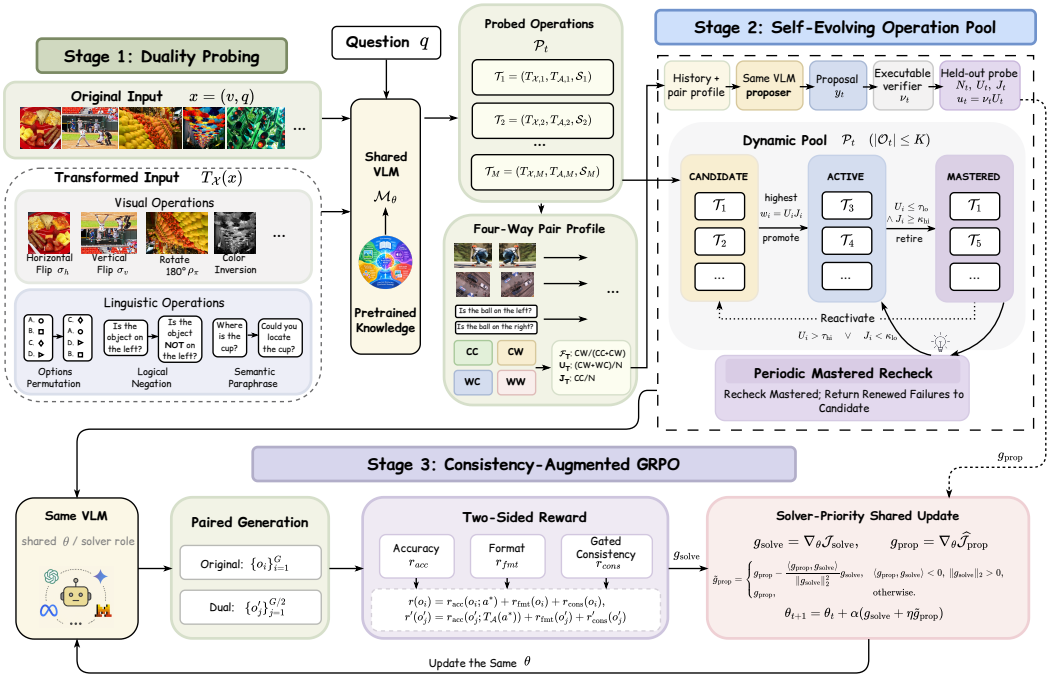

Figure 2: Overview of SAGE. Stage 1 probes $\mathcal{M}_\theta$ with duality operations to detect reasoning inconsistency. Stage 2 manages the operation pool through a promote–retire–recheck lifecycle. Stage 3 integrates duality consistency as an auxiliary reward within GRPO training.

2 METHOD

Robust spatial reasoning requires both correct answers and coherent predictions across inputs linked by known answer transformations. SAGE implements this principle in three stages: (1) *duality probing* detects inconsistencies on original–dual pairs; (2) a *self-evolving operation pool* autonomously discovers and schedules transformations; and (3) *consistency-augmented GRPO* converts these signals into training rewards. The overall workflow is illustrated in Figure 2.

2.1 STAGE 1: DUALITY PROBING

Contemporary VLMs achieve high accuracy on spatial reasoning benchmarks, yet correct answers alone do not establish that their predictions remain consistent under valid transformations. To formalize this, we introduce the notion of *duality complement*. Let $\mathcal{M}_\theta$ denote a VLM parameterized by θ that, given a visual input $v \in \mathcal{V}$ and a query $q \in \mathcal{Q}$, produces a distribution over answers $\mathcal{M}_\theta(a \mid v, q)$. Let $x = (v, q) \sim \mathcal{D}$ with $x \in \mathcal{X} = \mathcal{V} \times \mathcal{Q}$, and let $\mathcal{A}$ be the canonical answer space. We denote the decoded answer and ground truth by $\hat{a}_\theta(x)$ and $a^*(x) \in \mathcal{A}$, respectively, with $\perp \notin \mathcal{A}$ indicating extraction failure and counted as incorrect.

Definition 1 (Duality operation). *A duality operation is a triple $\mathcal{T} = (T_{\mathcal{X}}, T_{\mathcal{A}}, \mathcal{S})$, where $T_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ transforms the input, $T_{\mathcal{A}} : \mathcal{A} \to \mathcal{A}$ is the induced bijective answer transformation, and $\mathcal{S} \subseteq \mathcal{X}$ is the applicability domain, such that*

$$\forall x \in \mathcal{S} : \qquad a^*\big(T_{\mathcal{X}}(x)\big) = T_{\mathcal{A}}\big(a^*(x)\big). \tag{1}$$

Equation (1) defines answer-level equivariance through a deterministic relation between dual answers, with invariance as the special case $T_{\mathcal{A}} = \mathrm{id}$.

Definition 2 (Duality consistency). *For $\Pr_{x\sim\mathcal{D}}[x \in \mathcal{S}] > 0$, let $\mathcal{D}_{\mathcal{S}} = \mathcal{D}(\cdot \mid x \in \mathcal{S})$. The duality consistency of $\mathcal{M}_\theta$ with respect to $\mathcal{T}$ is*

$$\mathcal{C}_{\mathcal{T}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{S}}}\left[\mathbf{1}\left[T_{\mathcal{A}}(\hat{a}_\theta(x)) = \hat{a}_\theta(T_{\mathcal{X}}(x))\right]\right], \tag{2}$$

where the indicator is defined as zero if either answer extraction fails.

Since consistent answers can both be wrong, we define the hidden dual failure rate conditioned on original correctness, requiring $\Pr_{x\sim\mathcal{D}_\mathcal{S}}[\hat{a}_\theta(x) = a^*(x)] > 0$:

$$\mathcal{F}_\mathcal{T}(\theta) = \Pr_{x\sim\mathcal{D}_\mathcal{S}} \left[\hat{a}_\theta(T_\mathcal{X}(x)) \neq T_\mathcal{A}(a^*(x)) \,|\, \hat{a}_\theta(x) = a^*(x)\right]. \tag{3}$$

For a finer analysis, we partition applicable pairs into $\{\mathrm{CC}, \mathrm{CW}, \mathrm{WC}, \mathrm{WW}\}$, where C/W denote correct/wrong predictions on the original and dual inputs, respectively. Let $N^z_\mathcal{T}$ denote the count of case z, and let $N_\mathcal{T}$ be the total number of applicable pairs. We report the one-sided failure rate and joint-correctness rate:

$$U_\mathcal{T} = \frac{N^{\mathrm{CW}}_\mathcal{T} + N^{\mathrm{WC}}_\mathcal{T}}{N_\mathcal{T}}, \qquad J_\mathcal{T} = \frac{N^{\mathrm{CC}}_\mathcal{T}}{N_\mathcal{T}}. \tag{4}$$

Visual duality. These operations transform v while retaining q. Initial operations include *horizontal reflection* σ_h and *vertical reflection* σ_v, whose answer maps swap left↔right and top↔bottom, respectively; $180°$ *rotation* ρ_π, which composes both mappings; and *appearance perturbations* (color inversion, grayscale), with $T_\mathcal{A} = \mathrm{id}$ only when the answer is unaffected. Flips exclude text-reading questions and samples for which the directional mapping is insufficient. Input transformations need not be invertible, but must satisfy Eq. (1) on their applicability domain.

Linguistic duality. These operations retain v and transform q through *option permutation*, which reorders C choices by $\pi \in S_C$ with $T_\mathcal{A}(a) = \pi^{-1}(a)$; *logical negation*, which complements an unambiguous binary answer and applies to multiple-choice questions only when the transformed answer is unique; or *semantic paraphrase*, where $T_\mathcal{X}(v, q) = (v, q')$ preserves meaning and $T_\mathcal{A} = \mathrm{id}$. Additional visual or linguistic operations can be discovered subject to Eq. (1) on their applicability domains.

Composition. Valid $\mathcal{T}_1$ and $\mathcal{T}_2$ compose through $T_{\mathcal{X},1} \circ T_{\mathcal{X},2}$ and $T_{\mathcal{A},1} \circ T_{\mathcal{A},2}$, with applicability domain $\mathcal{S}_2 \cap T^{-1}_{\mathcal{X},2}(\mathcal{S}_1)$.

Efficient batch probing. We process N sampled applicable pairs in one batched inference through $\mathcal{M}_\theta$, recording all four outcomes, including original-side errors.

2.2 Stage 2: Self-Evolving Operation Pool

Not all duality operations are equally informative throughout training. Operations that the model already solves correctly on both sides provide limited additional supervision, whereas operations yielding consistently wrong predictions provide little useful consistency signal. We therefore maintain a dynamic pool $\mathcal{P}_t$ and allow the same trainable VLM to discover, evaluate, and solve new operations online.

Proposer-format initialization. A short SFT stage teaches the model to serialize the $(T_\mathcal{X}, T_\mathcal{A}, \mathcal{S})$ contract from context–proposal examples (h, y):

$$\mathcal{L}_{\text{prop-SFT}}(\theta) = -\mathbb{E}_{(h,y)} \sum_{\ell=1}^{|y|} \log \mathcal{M}_\theta(y_\ell \mid h, y_{<\ell}). \tag{5}$$

The resulting checkpoint $\mathcal{M}_{\theta_0}$ serves as both proposer and solver during online training; no separate proposer or learned judge is introduced.

Candidate discovery. Every E_p steps, the current model receives a proposer context h_t containing compact operation history and pair profiles, and samples a serialized proposal $y_t = (y_{t,1}, \ldots, y_{t,L})$. The proposal is parsed into a candidate $\widetilde{\mathcal{T}}_t = (\widetilde{T}_{\mathcal{X},t}, \widetilde{T}_{\mathcal{A},t}, \widetilde{\mathcal{S}}_t)$ containing executable input- and answer-transformation rules, an applicability predicate, and a rationale. Applying the rule to labeled source items yields original–dual pairs with mechanically derived answers. The initial validated operations provide a starting set, which the proposer continually expands or refines.

Validation and admission. A parameter-free verifier returns $\nu_t \in \{0, 1\}$ through executable checks, not a proof of validity. Passing candidates receive probe scores (N_t, U_t, J_t) on a buffer reserved from solver rollouts for training feedback, not independent evaluation. Proposal utility is

$$u_t = \nu_t U_t, \tag{6}$$

with $U_t = 0$ for failed validation or empty probes. A proposal enters $\mathcal{P}_t$ as CANDIDATE only if $\nu_t = 1$, $N_t \geq N_{\min} > 0$, and $U_t \geq \tau_{\text{add}} \in (0, 1]$.

Proposal update. Because y_t is discrete, we optimize its likelihood with the sample policy-gradient surrogate

$$\widehat{\mathcal{J}}_{\text{prop},t}(\theta) = \text{sg}(u_t - b_t) \sum_{\ell=1}^{L} \log \mathcal{M}_\theta(y_{t,\ell} \mid h_t, y_{t,<\ell}), \tag{7}$$

where sg stops gradients through the utility and baseline. Starting from $b_0 = 0$, the baseline is updated after each proposal as $b_{t+1} = \rho b_t + (1 - \rho)u_t$, $0 \leq \rho < 1$; the surrogate contributes a positive update when $u_t > b_t$.

States and transitions. Each operation $\mathcal{T}_i$ stores exponentially smoothed scores U_i and J_i, initialized from its admission probe. Profiles are updated only with at least $N_{\min}$ applicable pairs; unsupported estimates remain unchanged. Each operation has state $s_i \in \{\text{CANDIDATE}, \text{ACTIVE}, \text{MASTERED}\}$, with active subset $\mathcal{O}_t \subseteq \mathcal{P}_t$ satisfying $|\mathcal{O}_t| \leq K$. We assign each candidate a priority

$$w_i = U_i J_i, \tag{8}$$

which favors operations that expose one-sided failures while retaining substantial joint correctness. Every E solver steps, supported probes trigger state transitions, while candidates fill vacant active slots:

$$s_i : \text{ACTIVE} \xrightarrow{U_i \leq \tau_{\text{lo}} \,\wedge\, J_i \geq \kappa_{\text{hi}}} \text{MASTERED}, \tag{9}$$

$$s_i : \text{MASTERED} \xrightarrow{U_i > \tau_{\text{hi}} \,\vee\, J_i < \kappa_{\text{lo}}} \text{CANDIDATE}, \tag{10}$$

$$s_i : \text{CANDIDATE} \xrightarrow{\text{highest } w_i,\ |\mathcal{O}_t| < K} \text{ACTIVE}. \tag{11}$$

Here τ_{lo}, τ_{hi}, κ_{lo}, and κ_{hi} control retirement and reactivation.

2.3 STAGE 3: CONSISTENCY-AUGMENTED GRPO

We integrate duality consistency into the GRPO framework (Shao et al., 2024). At each solver step, given $x = (v, q)$ and its label $a^* = a^*(x)$:

Generation. The model generates G completions $\{o_i\}_{i=1}^{G} \sim \mathcal{M}_{\theta_t}(\cdot \mid x)$, where $G \geq 4$ is even. An active operation $\mathcal{T} \in \mathcal{O}_t$ with $x \in \mathcal{S}_\mathcal{T}$ is sampled with probability proportional to its priority; if all applicable priorities are zero, sampling is uniform. The same model generates $G/2$ dual completions $\{o'_j\}_{j=1}^{G/2} \sim \mathcal{M}_{\theta_t}(\cdot \mid T_\mathcal{X}(x))$. If no active operation applies, the step uses original-only GRPO.

Reward. Rewards combine task accuracy, format adherence, and duality consistency, supervising original and dual completions against a^* and the mechanically derived $T_\mathcal{A}(a^*)$, respectively:

$$\begin{aligned} r(o_i) &= r_{\text{acc}}(o_i; a^*) + r_{\text{fmt}}(o_i) + r_{\text{cons}}(o_i), \\ r'(o'_j) &= r_{\text{acc}}(o'_j; T_\mathcal{A}(a^*)) + r_{\text{fmt}}(o'_j) + r'_{\text{cons}}(o'_j). \end{aligned} \tag{12}$$

The consistency bonus is gated by task correctness. Let

$$c_{ij} = \mathbf{1}\left[T_\mathcal{A}(\text{ans}(o_i)) = \text{ans}(o'_j)\right], \quad g_{ij} = \mathbf{1}\left[\text{ans}(o_i) = a^* \,\vee \text{ans}(o'_j) = T_\mathcal{A}(a^*)\right]. \tag{13}$$

The symmetric consistency bonuses are

$$r_{\text{cons}}(o_i) = \lambda \frac{2}{G} \sum_{j=1}^{G/2} g_{ij} c_{ij}, \quad r'_{\text{cons}}(o'_j) = \lambda \frac{1}{G} \sum_{i=1}^{G} g_{ij} c_{ij}. \tag{14}$$

where $\text{ans}(\cdot)$ extracts a canonical answer and $\lambda \geq 0$ controls the weight. Extraction failure sets $c_{ij} = g_{ij} = 0$. For a valid operation, bijectivity implies $g_{ij} c_{ij} = 1$ if and only if both answers are correct; consistent but wrong pairs receive no bonus.

Policy update. Following GRPO, we compute group-relative advantages separately for the original and dual prompts:

$$\hat{A}_i = \frac{r(o_i) - \bar{r}}{\sigma_r + \epsilon}, \qquad \hat{A}'_j = \frac{r'(o'_j) - \bar{r}'}{\sigma_{r'} + \epsilon}, \tag{15}$$

where each group supplies its own mean and standard deviation, with $\epsilon > 0$. Holding rollouts and advantages fixed, we equally weight the two clipped GRPO policy surrogates and apply a single KL penalty:

$$\mathcal{J}_{\text{solve}}(\theta) = \tfrac{1}{2}\mathcal{J}^{\text{orig}}_{\text{GRPO}}(\theta) + \tfrac{1}{2}\mathcal{J}^{\text{dual}}_{\text{GRPO}}(\theta) - \beta D_{\text{KL}}[\mathcal{M}_\theta(\cdot \mid x) \| \mathcal{M}_{\theta_{\text{ref}}}(\cdot \mid x)], \tag{16}$$

where $\beta \geq 0$ controls the KL penalty against a frozen reference model, applied to the original prompt and omitted when no reference is used. All four correctness outcomes participate in task rewards; a group with identical rewards has zero advantages.

Solver-priority shared update. Because proposal generation and solving share θ, their gradients may conflict. At $\theta = \theta_t$, we compute

$$g_{\text{solve}} = \nabla_\theta \mathcal{J}_{\text{solve}}, \qquad g_{\text{prop}} = \nabla_\theta \widehat{\mathcal{J}}_{\text{prop},t}, \tag{17}$$

and project only when the two gradients conflict:

$$\widetilde{g}_{\text{prop}} = \begin{cases} g_{\text{prop}} - \dfrac{\langle g_{\text{prop}}, g_{\text{solve}} \rangle}{\|g_{\text{solve}}\|_2^2} g_{\text{solve}}, & \langle g_{\text{prop}}, g_{\text{solve}} \rangle < 0 \text{ and } \|g_{\text{solve}}\|_2 > 0, \\ g_{\text{prop}}, & \text{otherwise.} \end{cases} \tag{18}$$

The shared update, with step size $\alpha > 0$ and proposer weight $\eta \geq 0$, is

$$\theta_{t+1} = \theta_t + \alpha\big(g_{\text{solve}} + \eta \widetilde{g}_{\text{prop}}\big). \tag{19}$$

The projection gives $\langle g_{\text{solve}}, \widetilde{g}_{\text{prop}} \rangle \geq 0$, mitigating gradient conflict with one trainable parameter and optimizer state. This is a first-order guarantee for the sampled solver surrogate under raw gradients, not for task performance or updates with momentum, preconditioning, or weight decay.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate SAGE on six video understanding benchmarks: MVBench Li et al. (2024), TempCompass Liu et al. (2024), VideoMMMU Hu et al. (2025), Video-MME Fu et al. (2025), VSI-Bench Yang et al. (2025b), and MMVU Zhao et al. (2025c); and seven spatial reasoning benchmarks: CV-Bench Tong et al. (2024), SPAR-Bench Zhang et al. (2026a), ViewSpatial-Bench Li et al. (2025b), MMSI-Bench Yang et al. (2026), MindCube Yin et al. (2025), OST-Bench Lin et al. (2026), and OmniSpatial Jia et al. (2026).

Baselines. For video understanding, we compare against a broad range of state-of-the-art VLMs, including proprietary models such as GPT-4o Hurst et al. (2024), Gemini-2.0-Flash Team et al. (2024), GPT-5, GPT-5.4 Singh et al. (2025), and Gemini-3-Pro-Preview, as well as open-source models including InternVL-2.5 Chen et al. (2025b), LLaMA-VID Li et al. (2025c), VideoLLaMA2 Cheng et al. (2024b), LongVA-7B Zhang et al. (2025a), VILA-1.5-40B Lin et al. (2024), Video-UTR-7B Yu et al. (2025), LLaVA-OneVision-7B Li et al. (2025a), Kangaroo-8B Liu et al. (2026a), Video-R1-7B Feng et al. (2025), Video-RTS-7B Wang et al. (2025c), VideoRFT-7B Wang et al. (2025b), ViSS-R1-7B Fang et al. (2025), Video-RAISE-7B Xi et al. (2026), Video-KTR-7B Wang et al. (2026), VideoSSR-8B He et al. (2026b), VideoAuto-R1-8B Liu et al. (2026b), and OneThinker-8B Feng et al. (2026). For spatial reasoning, we compare with GPT-4o, InternVL-2.5-8B, LLaVA-OneVision-7B, SpaceR Ouyang et al. (2025), VILASR Wu et al. (2026), Spatial-MLLM Wu et al. (2025a), SpatialLadder-3B Li et al. (2026b), VST-7B-RL Yang et al. (2025c), and SenseNova-SI Cai et al. (2026). Due to the computational cost of reproducing all baselines, we combine reported results with our own evaluations, following the settings specified in the original papers. Appendix D.2 details the result sources and evaluation protocols.

Table 1: Performance comparison on video understanding benchmarks.

Model	VSI-Bench		VideoMMMU		MMVU (MC)		MVBench		TempCmps		VideoMME	
Proprietary Models												
GPT-4o	34.0		61.2		75.4		-		-		71.9	
Gemini-2.0-Flash	45.4		-		-		-		-		-	
GPT-5	55.0		84.6		82.6		74.1		-		83.3	
GPT-5.4	-		87.1		-		78.3		-		-	
Gemini-3-Pro-Preview	52.5		87.6		-		-		-		-	
Qwen2.5-VL-based Models												
Video-R1-7B	34.6		49.8		64.2		62.7		72.6		57.4	
Video-RTS-7B	33.5		52.7		66.4		62.7		74.9		63.0	
VideoRFT-7B	36.8		51.1		**68.5**		62.1		73.7		59.8	
ViSS-R1-7B	37.3		51.7		66.1		65.6		75.3		60.5	
Video-RAISE-7B	34.8		52.8		66.4		64.3		74.7		57.4	
Video-KTR-7B	37.5		51.3		65.7		64.8		73.3		57.3	
Qwen2.5-VL-7B-Instruct	32.5		46.1		59.2		59.1		70.4		50.6	
+SAGE	**37.9**	(+5.4)	**53.4**	(+7.3)	65.8	(+6.6)	**68.1**	(+9.0)	**78.4**	(+8.0)	**66.5**	(+15.9)
Qwen3-VL-based Models												
VideoSSR-8B	40.3		63.1		71.2		68.9		**82.9**		67.6	
VideoAuto-R1-8B	42.0		65.0		71.0		71.8		80.6		67.9	
OneThinker-8B	41.0		66.2		70.5		69.4		80.1		66.5	
Qwen3-VL-4B-Instruct	44.2		53.0		62.7		58.3		71.6		51.6	
+SAGE	45.8	(+1.6)	63.2	(+10.2)	71.2	(+8.5)	68.5	(+10.2)	78.7	(+7.1)	67.9	(+16.3)
Qwen3-VL-8B-Instruct	41.3		57.8		66.7		60.3		75.6		54.3	
+SAGE	**48.0**	(+6.7)	**69.6**	(+11.8)	**75.4**	(+8.7)	**74.8**	(+14.5)	82.6	(+7.0)	**70.0**	(+15.7)

Training and evaluation. We conduct experiments with Qwen2.5-VL-7B-Instruct Bai et al. (2025b), Qwen3-VL-4B/8B-Instruct Bai et al. (2025a). We perform proposer SFT on DeepSeek-V4-Pro-generated duality code, followed by GRPO on 5K Video-R1 Feng et al. (2025) and 5K SpatialLadder Li et al. (2026b) samples. RL training uses 16 NVIDIA A100 GPUs with DeepSpeed ZeRO-3, a learning rate of 10^{-6}, batch size 1 per device, and 8 original rollouts per prompt for one epoch. Unless otherwise specified, $\lambda = 0.3$, $\beta = 0.04$, $K = 3$, $E = 100$, $E_p = 100$, $\eta = 0.01$, and $p_f = 0.2$. We use the same decoding and answer extraction procedure across methods and evaluate duality consistency on valid, applicable original–transformed pairs, reporting four correctness outcomes: correct–correct (CC), correct–wrong (CW), wrong–correct (WC), and wrong–wrong (WW). Detailed settings are provided in Appendix D.3.

3.2 Main Results

Video understanding. Table 1 shows that SAGE improves all three backbones, with the largest gains on VideoMME (+15.7–16.3 points) and substantial gains on MVBench (+9.0–14.5). By encouraging correct answers across original and transformed inputs under the prescribed answer mapping, duality training couples augmentation with a reasoning constraint. This may reduce reliance on orientation-specific or superficial cues, supporting more robust grounding across diverse scenes and events. The smaller gains on VSI-Bench (+1.6–6.7) suggest that transformation robustness alone does not resolve spatial-memory and metric-estimation demands, while VideoMMMU and MMVU additionally require knowledge acquisition or domain expertise. These results are consistent with improved generalization through robust visual reasoning, with benefits depending on each task's underlying bottlenecks.

Spatial reasoning. Table 2 shows consistent improvements across all seven benchmarks and three backbones, with larger gains on more demanding spatial reasoning tasks. CV-Bench shows relatively modest improvements, as the baselines are already strong on basic spatial judgments, while SPAR, ViewSpatial, MMSI, and OmniSpatial exhibit more substantial gains on structured, viewpoint-dependent, and multi-image reasoning. For Qwen3, the gains on MindCube and OST exceed 20 points, highlighting the benefit of the proposed training for integrating spatial evidence

Table 2: Performance comparison on spatial reasoning benchmarks.

Model	CV-Bench	SPAR	ViewSp.	MMSI	MindCube	OST	OmniSp.
Proprietary Models							
GPT-4o	75.4	36.4	32.6	30.3	38.8	-	-
Other Open-source Backbones							
InternVL-2.5-8B	76.5	36.3	43.2	25.7	18.7	20.4	24.3
LLaVA-OV-7B	58.3	31.2	27.5	24.5	47.3	19.5	23.1
SenseNova-SI-8B	82.4	42.1	**54.7**	**43.3**	47.6	43.8	27.5
Qwen2.5-VL-based Models							
SpatialLadder-3B	73.7	34.4	44.2	29.2	43.4	23.9	30.8
VST-7B-RL	86.5	32.8	43.5	34.8	44.9	38.0	32.2
Qwen2.5-VL-7B-Instruct	74.2	21.8	37.4	26.4	29.5	38.1	23.7
+SAGE	77.2 (+3.0)	34.4 (+12.6)	44.1 (+6.7)	32.3 (+5.9)	44.5 (+15.0)	46.3 (+8.2)	38.3 (+14.6)
Qwen3-VL-based Models							
Qwen3-VL-4B-Instruct	81.8	35.4	41.3	24.8	18.0	25.2	30.1
+SAGE	83.5 (+1.7)	41.4 (+6.0)	47.3 (+6.0)	31.9 (+7.1)	39.7 (+21.7)	**46.4** (+21.2)	39.2 (+9.1)
Qwen3-VL-8B-Instruct	82.9	37.2	42.0	26.2	26.4	25.7	32.9
+SAGE	**86.7** (+3.8)	**43.2** (+6.0)	46.2 (+4.2)	34.8 (+8.6)	**48.4** (+22.0)	46.1 (+20.4)	**40.4** (+7.5)

Table 3: Ablation study of duality supervision, operation design, self-evolution, and shared-model optimization.

Variant	Average Score	Consistency	Joint Correctness	Wrong-consistent	GPU-hours	Peak Memory (GiB/GPU)
Base	43.8	61.4	24.1	37.3	-	-
SAGE	52.9	75.9	45.7	30.2	26.8	76.2
A. Reward design (fixed paired stream)						
Original-only GRPO	47.7	62.3	28.8	33.5	18.7	73.1
Dual augmentation only	51.4	68.9	35.3	33.6	37.6	76.2
Ungated consistency	49.4	81.1	38.4	42.7	26.4	76.2
B. Operation families						
Visual only	48.8	74.8	39.3	35.5	23.4	77.0
Linguistic only	45.1	73.5	35.6	37.9	22.5	76.2
C. Operation selection and discovery						
Fixed-all	49.3	76.6	40.3	36.3	42.8	76.2
Random-K	47.0	74.1	37.7	36.4	26.6	76.1
Adaptive-K	49.8	76.8	41.0	35.8	27.4	76.2
D. Online evolution						
w/o online discovery	44.6	74.0	35.7	38.3	21.6	76.2
w/o retirement	48.6	77.1	39.0	38.1	28.8	76.2
w/o reactivation	50.2	73.8	41.5	32.3	26.9	76.2
E. Shared-model optimization						
Separate proposer–solver	52.9	76.3	45.9	30.4	39.0	79.8
Shared, w/o projection	51.4	74.7	42.9	31.8	27.1	76.2

across multiple views or observations. Overall, duality training improves not only basic spatial recognition but also more complex relational reasoning. Notably, the SAGE-trained 4B model consistently outperforms the untuned 8B model across all seven benchmarks, demonstrating the benefit of targeted relational supervision beyond model scaling.

3.3 Ablation Study

The results in Table 3 show that ordinary GRPO yields only limited gains, while its low semantic consistency suggests that the model does not reliably capture the underlying relations, limiting

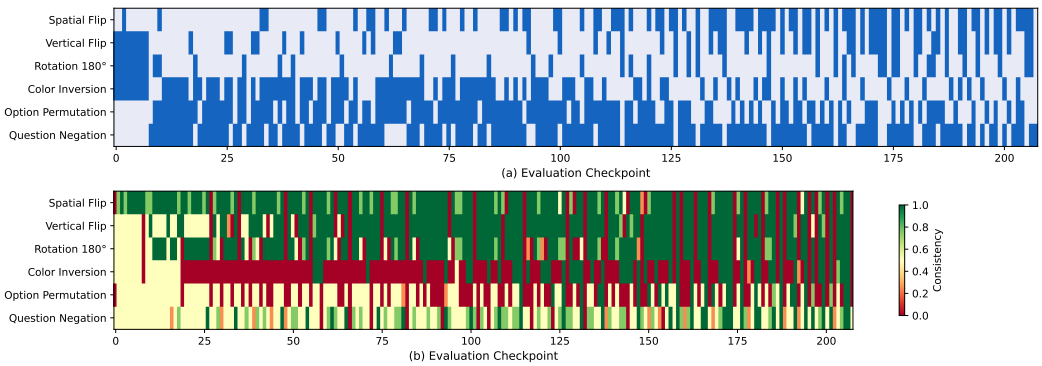

Figure 3: Operation lifecycle during SAGE training. Operations are proposed, promoted to the active pool, trained until mastery, and periodically re-evaluated for forgetting.

the resulting task improvement. Dual augmentation is competitive, yet updating separately on augmented inputs incurs additional optimizer steps; Appendix F.5 identifies conditions under which a joint update can avoid both this extra update cost and the order-dependent bias introduced by sequential updates. Ungated consistency achieves the highest consistency largely through wrong-consistent predictions, highlighting that agreement alone is not reliable supervision. Combining visual and linguistic operations also outperforms either family alone, suggesting complementary supervision across changes in visual presentation and question formulation. For operation scheduling, both Fixed-all and Random-K underperform SAGE, while Adaptive-K simply allocates training opportunities by operation counts without probing their utility, also falling short of SAGE; removing discovery, retirement, or reactivation further reduces performance, supporting their complementary roles in expanding coverage, reallocating training effort, and revisiting forgotten relations. Separate proposer–solver models achieve similar accuracy but duplicate trainable parameters and optimizer states, increasing GPU-hours and memory. Removing gradient projection also degrades performance, indicating that gradient decomposition helps mitigate interference and enables effective role sharing within a single model.

3.4 Further Analysis

Operation lifecycle. Figure 3 shows the dynamics of selected operations. Mastered operations yield active slots to candidates exposing unresolved inconsistencies, forming an adaptive curriculum that focuses training on remaining reasoning weaknesses.

Correctness and generalization. Joint correctness rises by 21.6 points and wrong-consistent predictions fall by 7.1 points; a 20.8-point gain in unseen joint correctness over Base suggests transfer beyond the training pool. Appendix E.3 details both evaluations.

Proposal quality and sensitivity. Generated proposals are 88.33% valid, 70.00% novel, and 56.67% useful, extending supervision beyond the initial pool. Gains over Base persist across six λ values and five K settings; Appendix E.3 reports proposal quality and sensitivity.

4 Conclusion

We presented SAGE, a self-evolving framework that strengthens spatial reasoning in VLMs through geometric and linguistic duality. By discovering informative transformations and enforcing their answer relations, SAGE encourages reasoning grounded in spatial structure rather than surface patterns. Experiments across six video understanding and seven spatial reasoning benchmarks demonstrate improvements across multiple backbones, supporting transformation-aware supervision as a complementary signal for post-training. A shared proposer–solver model integrates operation discovery and reasoning within a unified learning process, with solver-priority updates mitigating interference between these roles. Future work includes extending duality to richer 3D and multi-view transformations and exploring self-evolving relational supervision beyond spatial reasoning.

AI USE STATEMENT

LLM-based tools assisted with writing, coding, and proof drafting. DeepSeek-V4-Pro was used to prepare proposer-format SFT examples (Section 3.1). The authors are responsible for verifying AI-assisted material and for the final content of this work.

ETHICS STATEMENT

SAGE studies spatial reasoning using existing datasets and pretrained models. Improved consistency does not guarantee factual correctness or safe behavior, and inherited biases and privacy risks may persist. Generated transformation code requires restricted execution and validation. Applications in robotics, navigation, or surveillance require additional safety and privacy assessment; benchmark gains alone do not establish deployment readiness. Use and redistribution of datasets, model weights, and images must respect their licenses and applicable restrictions.

REPRODUCIBILITY STATEMENT

Given resource constraints, we plan to release the training data, code, and model weights upon acceptance. The paper and appendices document the training configuration, hyperparameters, evaluation protocols, and baseline result sources to facilitate reproduction.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, June 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=k7vcuqLK4X.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093. Curran Associates, Inc., 2024a. doi: 10.52202/079017-4293. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f38cb4cf9a5eaa92b3cfa481832719c6-Paper-Conference.pdf.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024b. URL https://arxiv.org/abs/2406.07476.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Gyouk Chu, Myeongho Jeon, and Eunho Yang. J-zero: Unified challenger–solver–judge co-evolution from zero data, 2026. URL https://arxiv.org/abs/2608.26582.

Bo Fang, Yuxin Song, Qiangqiang Wu, Haoyuan Sun, Wenhao Wu, and Antoni B. Chan. Viss-r1: Self-supervised reinforcement video reasoning, 2025. URL https://arxiv.org/abs/2511.13054.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=a2JTVVvcEl.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5432–5443, June 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26274–26284, June 2026a.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Videossr: Video self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26530–26540, June 2026b.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. URL https://arxiv.org/abs/2501.13826.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 130770–130790, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d49b9aacebda61051166335af6fd3061-Paper-Conference.pdf.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6nZKT2rL0H.

Sai Srinivas Kancheti, Aditya Kanade, Rohit Sinha, Vineeth N Balasubramanian, and Tanuja Ganu. Faithful grpo: Improving visual spatial reasoning in multimodal language models via constrained policy optimization, 2026. URL https://arxiv.org/abs/2604.08476.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32803–32827, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1514. URL https://aclanthology.org/2026.acl-long.1514/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Dinging Li, Yingxiu Zhao, Xinrui Cheng, Kangheng Lin, Hongbo Peng, Hongxing Li, Zixuan Wang, Yuhong Dai, Haodong Li, Jia Wang, Yukang Shi, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Spatialevo: Self-evolving spatial intelligence via deterministic geometric environments, 2026a. URL https://arxiv.org/abs/2604.14144.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025b. URL https://arxiv.org/abs/2505.21500.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=KtrFXlvgrK.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, June 2024.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 323–340, Cham, 2025c. Springer Nature Switzerland. ISBN 978-3-031-72952-2.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Multi-view spatial transformation reasoning by reinforcing multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12041–12051, June 2026c.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26689–26699, June 2024.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-bench: Evaluating the capabilities of MLLMs in online spatio-temporal scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=vAkVKIOtcN.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 06 2023a. ISSN 2307-387X. doi: 10.1162/tacl_a_00566. URL https://doi.org/10.1162/tacl_a_00566.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *International Journal of Computer Vision*, 134(3):114, feb 2026a. ISSN 1573-1405. doi: 10.1007/s11263-025-02620-2. URL https://doi.org/10.1007/s11263-025-02620-2.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, Ernie Chang, Saksham Suri, Hongyu Xu, Qi Qian, Wei Wen, Balakrishnan Varadarajan, Zhuang Liu, Hu Xu, Florian Bordes, Raghuraman Krishnamoorthi, Bernard Ghanem, Vikas Chandra, and Yunyang Xiong. Videoauto-r1: Video auto reasoning via thinking once, answering twice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32829–32839, June 2026b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning, 2025. URL https://arxiv.org/abs/2504.01805.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in VLMs for the metaverse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdQzLC0Zra.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N0I2RtD8je.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169, June 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Henry Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in

vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XpL5eqjCjF.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356. Curran Associates, Inc., 2024. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=4oYxzssbVg.

Peiyao Wang and Haibin Ling. SVQA-r1: Reinforcing spatial reasoning in MLLMs via view-consistent reward optimization. In *The First Workshop on Efficient Spatial Reasoning*, 2026. URL https://openreview.net/forum?id=o2E8oa2frj.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=3pORFyKzh1.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28126–28140, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1428. URL https://aclanthology.org/2025.emnlp-main.1428/.

Ziyue Wang, Sheng JIn, ZHONGRONG ZUO, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Jiang Xudong. Video-ktr: Reinforcing video reasoning via key token attribution. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 71553–71570, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/749b64078a64fa5734a49fb40bc9fd65-Paper-Conference.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=RnXS7aK4rK.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=yyWeSAsOhs.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=yyWeSAsOhs.

Xing Xi, Yu Qiu, Ronghua Luo, Peixian Chen, and peilin tong. Let vlms grade their own thoughts: A self-quantification approach to reasoning-aware reward modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26487–26496, June 2026.

Fan Yang, Yousong Zhu, Xin Li, Yufei Zhan, Hongyin Zhao, Shurong Zheng, Yaowei Wang, Ming Tang, and Jinqiao Wang. FOCUS: Unified vision-language modeling for interactive editing driven by referential segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=FACJ0478oQ.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, June 2025b.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning, 2025c. URL https://arxiv.org/abs/2511.05491.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, June 2025d.

Senqiao Yang, Junyi Li, Xin Lai, Jinming Wu, Wei Li, Zejun MA, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Visionthink: Smart and efficient vision language model via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025e. URL https://openreview.net/forum?id=R6m6bNnmWm.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-bench: A benchmark for multi-image spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gHRoX4vXm3.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025. URL https://openreview.net/forum?id=HBllcu0rmi.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and Wenbing Tao. Unhackable temporal rewarding for scalable video mllms, 2025. URL https://arxiv.org/abs/2502.12081.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 110935–110971. Curran Associates, Inc., 2024. doi: 10.52202/079017-3522. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c848b7d3adc08fcd0bf1df3101ba6728-Paper-Conference.pdf.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026a. URL https://openreview.net/forum?id=GzgPleFl8f.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. doi: 10.1109/TPAMI.2024.3369699.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=30RAWQVGlx.

Xinyu Zhang, Yuxuan Dong, Lingling Zhang, Chengyou Jia, Zhuohang Dang, Basura Fernando, Jun Liu, and Mike Zheng Shou. CoFFT: Chain of foresight-focus thought for visual language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=AQnjBIFCBQ.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and DING ZHAO. Spin-bench: Perspective and rotation as a lens on spatial reasoning in vlms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 70072–70141, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/724be4472168f31ba1c9ac630f15dec8-Paper-Conference.pdf.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=neZSGqhxDa.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11071–11080, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755703. URL https://doi.org/10.1145/3746027.3755703.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489, June 2025c.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13624–13634, June 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1. URL https://doi.org/10.1007/s11263-022-01653-1.

Mike Zhou, Fenil Bardoliya, Vivek Gupta, and Dan Roth. Program-of-thought reveals LLM abstraction ceilings. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4911–4919, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.257. URL https://aclanthology.org/2026.findings-eacl.257/.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):13782–13790, Mar. 2026b. doi: 10.1609/aaai.v40i16.38386. URL https://ojs.aaai.org/index.php/AAAI/article/view/38386.

A Related Work

A.1 Vision Language Models

The rapid evolution of VLMs has significantly transformed multimodal perception Zhang et al. (2023); Maaz et al. (2024); Zhang et al. (2024); Shu et al. (2025). Foundational frameworks like CLIP Radford et al. (2021) and BLIP Li et al. (2022) established cross-modal alignment, while subsequent architectures such as LLaVA Liu et al. (2023b), Flamingo Alayrac et al. (2022), and Qwen3-VL Bai et al. (2025a) advanced the field by integrating visual encoders with LLMs for sophisticated reasoning. Beyond architectural scaling, innovative strategies have emerged to refine model efficiency and perception. For instance, CoOp Zhou et al. (2022) introduced learnable context optimization, while VisionZip Yang et al. (2025d) and VisionThink Yang et al. (2025e) improve inference efficiency through visual token compression and reinforcement learning-based resolution scaling, respectively. Furthermore, recent works like CoFFT Zhang et al. (2025b) and FOCUS Yang et al. (2025a) have pushed the boundaries of visual reasoning and interactive editing by incorporating iterative foresight-focus decoding and unified segmentation-aware perception. Despite these advancements, existing VLMs often prioritize surface-level statistical patterns over genuine logical invariances Berglund et al. (2024). Consequently, they frequently exhibit reasoning inconsistencies and struggle to fully comprehend complex spatial and temporal geometries, particularly when faced with Out-Of-Distribution (OOD) visual arrangements Liu et al. (2023a).

A.2 Reinforcement Learning for Reasoning

The integration of Reinforcement Learning from Human Feedback (RLHF) Christiano et al. (2017); Stiennon et al. (2020); Ouyang et al. (2022) has become a standard paradigm for aligning large models, with algorithms like PPO Schulman et al. (2017) and DPO Rafailov et al. (2023) paving the way for goal-directed optimization. Building on these, Group Relative Policy Optimization (GRPO) Shao et al. (2024) has emerged as a more efficient alternative by computing advantages within sampled groups, thereby eliminating the need for a separate value network and simplifying the integration of custom reward signals. Early explorations utilized VLMs as zero-shot reward models to guide agent behavior Rocamonde et al. (2024), while Zhai et al. Zhai et al. (2024) leveraged RL to fine-tune VLMs as decision-making agents via Chain-of-Thought (CoT) Wei et al. (2022) exploration. More recently, VL-Rethinker Wang et al. (2025a) adapted GRPO to incentivize self-reflection in multimodal reasoning, and VAPO Tian et al. (2026) introduced vision-anchored rewards to mitigate the "visual forgetting" typically associated with prolonged reasoning. Our work similarly adopts the GRPO framework, as its flexible reward structure allows us to incorporate geometric and linguistic duality as self-correcting signals. By verifying consistency automatically across semantically equivalent transformations, SAGE enables a self-evolving training process that enhances spatial reasoning without the burden of extensive manual annotations.

A.3 Spatial reasoning in VLMs

Spatial reasoning remains a persistent challenge for vision-language models, as they often struggle with 3D geometries and relative object configurations Liu et al. (2023a); Chen et al. (2024; 2025a). Early efforts like SpatialRGPT Cheng et al. (2024a) and SpatialVLM Chen et al. (2024) addressed these limitations by introducing depth-aware plugins or internet-scale metric spatial data, while AdaptVis Chen et al. (2025a) explored inference-time attention sharpening to rectify misdirected focus. Recently, the R1-style reinforcement learning paradigm has been extended to spatial domains. Video-R1 Feng et al. (2025) utilized T-GRPO to incentivize temporal modeling, and ViLaSR Wu et al. (2025b) integrated interwoven thinking with visual drawing. Furthermore, frameworks like MetaSpatial Pan & Liu (2026) and SpatialLadder Li et al. (2026b) proposed physics-aware modulation and progressive training hierarchies to bridge the gap between perception and reasoning. More recently, SVQA-R1 Wang & Ling (2026) employed view-consistent rewards via static spatial perturbations, but its training signal remains static and cannot adapt to the model's evolving weaknesses. While Faithful GRPO Kancheti et al. (2026) and SpatialEvo Li et al. (2026a) introduce forms of self-correction or evolving environments, they are not designed to enforce transformation-level duality consistency. As a result, these methods fail to guarantee that the underlying spatial logic is truly mastered; a model might pass these specialized constraints yet still fail on fundamental dual transformations, leading to poor generalization on OOD tasks.

A.4 Self-evolving reasoning systems

Self-evolving reasoning has emerged as a promising approach to reducing dependence on fixed training curricula by adapting supervision to the learner's capabilities. R-Zero Huang et al. (2026) co-evolves separate Challenger and Solver models, while J-Zero Chu et al. (2026) introduces a co-adapting Judge to refine the evaluation signal. Complementary approaches ground self-improvement in executable environments or visual observations: Absolute Zero Zhao et al. (2025a) jointly trains a shared proposer–solver on code reasoning tasks, whereas VisPlay He et al. (2026a) develops visual reasoning through question generation over unlabeled images. Although zero-data self-evolution offers an elegant route to autonomous learning, task diversity can remain constrained by the generator's learned priors. R-Zero reports degradation over successive iterations and identifies diversity loss and bias amplification as possible contributors. Extending zero-data self-evolution to multimodal tasks may additionally require image or audio generators to create the observations themselves. The resulting dependence on generation quality and cross-modal alignment can limit training effectiveness, particularly when spatial reasoning requires precise geometric relations and reliable answer labels. SAGE instead grounds operation discovery in existing labeled images and videos, proposing geometric and linguistic transformations with explicit answer mappings. This anchors scene content and supervision to source examples while allowing the training curriculum to evolve with the model's weaknesses, without requiring a separate image generator or learned judge. By probing the same underlying relations across different presentations, SAGE is designed to discourage presentation-specific shortcuts and promote spatial generalization, rather than relying solely on diversity generated from the model's own knowledge.

B Additional Method Details

B.1 Operation Families and Composition

Executable validation. The parameter-free verifier checks parsing, restricted compilation, output type and shape, determinism, duplicates, the declared answer map, and applicability exclusions. For example, it rejects an identity answer map for a left/right flip. These checks provide operational acceptance, not a proof of semantic validity; a valid duality operation must still satisfy Eq. (1) and have bijective $T_{\mathcal{A}}$. Surviving candidates are evaluated using the pair counts and scores in Eq. (4).

Visual duality. The auditable seed operations include: (i) horizontal flip, for which $T_{\mathcal{A}}$ swaps left↔right; (ii) vertical flip, which swaps top↔bottom; (iii) 180° rotation, whose answer map composes the horizontal and vertical mappings; and (iv) appearance perturbations, for which $T_{\mathcal{A}} = \mathrm{id}$ only when the applicability predicate excludes appearance-dependent questions. Flips also exclude text reading and other orientation-sensitive samples unless the proposal supplies a valid answer map.

Linguistic duality. For option permutation, transformed position k contains original option $\pi(k)$ for $\pi \in S_C$, so $T_{\mathcal{A}}(a) = \pi^{-1}(a)$. Binary logical negation maps the answer to its complement and excludes ambiguous or non-binary questions. Semantic paraphrase uses $T_{\mathcal{A}} = \mathrm{id}$ and compares outputs with the task's deterministic answer extractor rather than raw strings.

Composition. If $\mathcal{T}_1 = (T_{\mathcal{X},1}, T_{\mathcal{A},1}, \mathcal{S}_1)$ and $\mathcal{T}_2 = (T_{\mathcal{X},2}, T_{\mathcal{A},2}, \mathcal{S}_2)$ satisfy Eq. (1), their composition uses $T_{\mathcal{X},1} \circ T_{\mathcal{X},2}$, $T_{\mathcal{A},1} \circ T_{\mathcal{A},2}$, and domain $\mathcal{S}_2 \cap T_{\mathcal{X},2}^{-1}(\mathcal{S}_1)$. This preserves the declared answer relation on the restricted domain, but executable and probe checks remain necessary because valid code does not prove semantic validity.

B.2 Training Iteration and Pool Updates

For a horizontal-reflection proposal, the verifier checks the input transformation, the left/right answer mapping, and the applicability predicate. The model is then evaluated on applicable original–dual pairs. Their four-way correctness profile determines the one-sided failure utility U and joint correctness J; a candidate is admitted only with sufficient probe support and utility.

For a question asking where a cup lies relative to a plate, a source answer of "left" maps to "right" after reflection. Both inputs receive task supervision. If only one answer is correct, the pair receives

Algorithm 1 Shared-Parameter Generative SAGE

Input: proposer-format-SFT policy $\mathcal{M}_{\theta_0}$; $\mathcal{D}, \mathcal{D}_{\text{probe}}, \mathcal{P}_{\text{seed}}$; T_{train}; even $G \geq 4$; E, E_p, M, K; $\lambda, \beta, \alpha, \eta, \rho, p_f, \epsilon$; $N_{\min}, \tau_{\text{add}}, \tau_{\text{lo}}, \tau_{\text{hi}}, \kappa_{\text{lo}}, \kappa_{\text{hi}}$
$\mathcal{P}_0 \leftarrow$ VALIDATESEEDS($\mathcal{P}_{\text{seed}}$); probe seeds and retain those with $N_i \geq N_{\min}$
Initialize profiles and states; set $b_0 \leftarrow 0$ and derive $\mathcal{O}_0$ via Eq. (11)
for step $t = 0, 1, \ldots, T_{\text{train}} - 1$ **do**
  $\widehat{g}_{\text{prop},t} \leftarrow 0$; $b_{t+1} \leftarrow b_t$
  if $(t+1) \bmod E_p = 0$ **then**
    $h_t \leftarrow$ PROPOSERCONTEXT($\mathcal{P}_t$, PROFILE$_t$); sample $y_t \sim \mathcal{M}_{\theta_t}(\cdot \mid h_t)$
    $\widetilde{\mathcal{T}}_t \leftarrow$ PARSE(y_t); $\nu_t \leftarrow$ EXECUTABLECHECKS($\widetilde{\mathcal{T}}_t$)
    $(N_t, U_t, J_t) \leftarrow (0, 0, 0)$
    if $\nu_t = 1$ **then**
      $(\text{CC}, \text{CW}, \text{WC}, \text{WW}) \leftarrow$ FOURWAYPROBE($\mathcal{M}_{\theta_t}, \widetilde{\mathcal{T}}_t, \mathcal{D}_{\text{probe}}$)
      Compute N_t; if $N_t > 0$, compute U_t, J_t
    end if
    Compute u_t, $\widehat{g}_{\text{prop},t}$, and b_{t+1} via Eqs. (6)–(7)
    if $\nu_t = 1$ and $N_t \geq N_{\min}$ and $U_t \geq \tau_{\text{add}}$ **then**
      $\mathcal{P}_t \leftarrow$ BOUNDEDINSERT($\mathcal{P}_t, \widetilde{\mathcal{T}}_t$, CANDIDATE, M) with its admission profile
    end if
  end if
  Sample $(x, a^*) \sim \mathcal{D}$ and generate $\{o_i\}_{i=1}^{G} \sim \mathcal{M}_{\theta_t}(\cdot \mid x)$
  $\mathcal{O}_t^x \leftarrow \{\mathcal{T}_i = (T_{\mathcal{X},i}, T_{\mathcal{A},i}, \mathcal{S}_i) \in \mathcal{O}_t : x \in \mathcal{S}_i\}$
  if $\mathcal{O}_t^x \neq \emptyset$ **then**
    Sample $\mathcal{T}_i \in \mathcal{O}_t^x$ proportional to w_i, uniformly if all priorities are zero
    $\{o'_j\}_{j=1}^{G/2} \sim \mathcal{M}_{\theta_t}(\cdot \mid T_{\mathcal{X},i}(x))$
    Compute $r(o_i)$ and $r'(o'_j)$ via Eqs. (12)–(14)
    Compute the two-sided $\mathcal{J}_{\text{solve}}$ and $g_{\text{solve}} = \nabla_\theta \mathcal{J}_{\text{solve}}$ via Eq. (16)
  else
    Set $r_{\text{cons}}(o_i) = 0$; compute original-only $\mathcal{J}_{\text{solve}}$ and g_{solve}
  end if
  Set $g_{\text{prop}} = \widehat{g}_{\text{prop},t}$; compute $\widetilde{g}_{\text{prop}}$ and update shared θ_{t+1} via Eqs. (18)–(19)
  if $(t+1) \bmod E = 0$ **then**
    Using $\mathcal{M}_{\theta_{t+1}}$, probe every operation in $\mathcal{O}_t$ and update its profile via Eq. (20)
    If any MASTERED operation exists, probe one with probability p_f and update its profile via Eq. (20)
    Apply Eqs. (9)–(10); fill vacant active slots with highest-priority candidates via Eq. (11)
  end if
  Set $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t$ and $\mathcal{O}_{t+1} \leftarrow \{\mathcal{T}_i \in \mathcal{P}_{t+1} : s_i = \text{ACTIVE}\}$
end for

no consistency bonus, but neither side is discarded from the task objective. Each prompt's advantages are normalized within its own rollout group. After the shared policy update, supported probes refresh the operation profiles using Eq. (20). Retirement requires both $U \leq \tau_{\text{lo}}$ and $J \geq \kappa_{\text{hi}}$; a mastered operation returns to the candidate pool when $U > \tau_{\text{hi}}$ or $J < \kappa_{\text{lo}}$.

C PSEUDOCODE

In this section, we present the pseudocode of SAGE, as shown in Algorithm 1.

D EXPERIMENTAL SETUP DETAILS

D.1 BENCHMARK DESCRIPTIONS

We evaluate SAGE on 13 benchmarks spanning video understanding and spatial reasoning. Below we describe each benchmark in detail, along with representative examples.

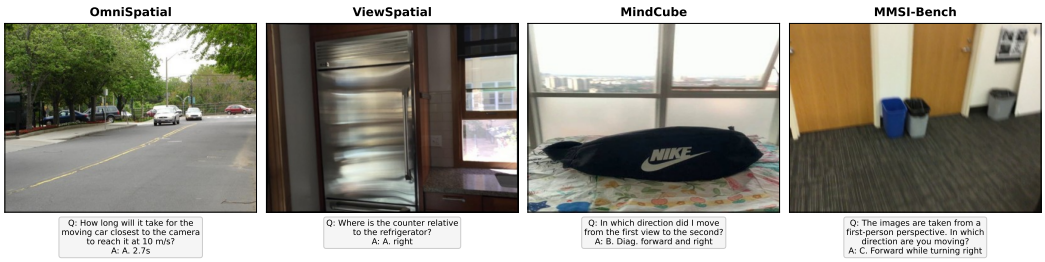

Figure 4: Representative examples from four spatial reasoning benchmarks. Each example shows the input image, the spatial question, and the ground-truth answer. These questions require understanding of object distances (OmniSpatial), relative positions (ViewSpatial), egomotion and perspective changes (MindCube), and first-person movement direction (MMSI-Bench).

D.1.1 Video Understanding Benchmarks

- **MVBench** Li et al. (2024) (4,000 samples). A comprehensive multi-modal video understanding benchmark covering 20 temporal understanding tasks, including action sequence recognition, scene transition detection, object tracking, and counterfactual inference. All questions are multiple-choice. *Example:* "What is the action performed by the person in the video?" with options describing different activities.
- **TempCompass** Liu et al. (2024) (7,540 samples). A temporal video description matching benchmark designed to test whether VLMs genuinely understand temporal dynamics or merely rely on static frame features. Questions probe fine-grained temporal perception such as speed, order, and direction of events.
- **VideoMME** Fu et al. (2025) (2,700 samples). A large-scale evaluation benchmark covering diverse video understanding scenarios across short ($<$2min), medium (4–15min), and long ($>$30min) videos, with questions spanning perception, cognition, and knowledge domains.
- **VideoMMMU** Hu et al. (2025) (900 samples). A challenging benchmark requiring domain-specific expert knowledge from video content, covering disciplines such as science, engineering, medicine, and humanities.
- **VSI-Bench** Yang et al. (2025b) (5,130 samples). A visual spatial intelligence benchmark constructed from 288 indoor video scenes. It contains both multiple-choice questions (2,490) and numerical regression tasks (2,640) across 10 sub-categories: object counting, absolute distance estimation, relative distance comparison, object size estimation, room size estimation, relative direction (easy/medium/hard), appearance order, and route planning. Following the original benchmark protocol, we report the average score across all 10 sub-tasks, using accuracy for multiple-choice and mean relative accuracy (MRA) for regression tasks. *Example:* "If I am standing by the refrigerator and facing the table, is the dishwasher to my front-left, front-right, back-left, or back-right?"
- **MMVU** Zhao et al. (2025c) (625 samples). A multi-discipline video understanding benchmark with questions from domains including science, healthcare, engineering, and daily life, requiring both visual understanding and domain knowledge.

D.1.2 Spatial Reasoning Benchmarks

- **CV-Bench** Tong et al. (2024) (2,638 samples). Transforms classic computer vision tasks into VQA format, with a 2D subset (1,438 samples) evaluating spatial relationships and object counting, and a 3D subset (1,200 samples) evaluating depth ordering and relative distance. Images are sourced from ADE20K, COCO, and OMNI3D with manually verified annotations.
- **SPAR-Bench** Zhang et al. (2026a) (7,207 samples). A large-scale 3D spatial perception and reasoning benchmark spanning 33 spatial tasks across single-view, multi-view, and video settings, sourced from 4,500 real-world scenes. Tasks include distance estimation, object-to-object spatial relations, spatial imagination, and route planning.

- **ViewSpatial-Bench** Li et al. (2025b) (5,712 samples). Evaluates multi-perspective spatial localization by requiring models to reason about the relative positions of objects from different camera viewpoints. Questions ask about directional relationships (e.g., left, right, front, back) between objects in 3D scenes. *Example:* "Could you tell me the location of the counter in comparison to the refrigerator?" with options: A. right, B. front-up, C. back-left, D. front.
- **MMSI-Bench** Yang et al. (2026) (1,000 samples). A human-annotated benchmark for multi-image spatial intelligence, where each question involves two or more images from real-world scenes including autonomous driving, robotic manipulation, and indoor scanning. Covers six types of positional relationships, two types of attribute reasoning, two types of motion reasoning, and multi-step reasoning. *Example:* "The images are taken continuously from a first-person perspective. In which direction are you moving?"
- **MindCube** Yin et al. (2025) (21,154 samples). Evaluates spatial mental modeling from limited visual views with questions testing cognitive mapping, perspective-taking, and mental simulation across 3,268 images. The model must form a coherent spatial mental model from partial observations to answer questions about unseen perspectives. *Example:* "Based on these two views showing the same scene: in which direction did I move from the first view to the second view?"
- **OST-Bench** Lin et al. (2026) (10,165 QA turns across 1.4K scenes; we evaluate 3,194 questions reformulated as independent single-turn instances). Tests online spatio-temporal scene understanding by requiring models to track object existence, positions, and state changes across sequential observations in 3D environments. Questions probe whether models can maintain a consistent internal representation of the evolving scene.
- **OmniSpatial** Jia et al. (2026) (8,400+ QA pairs in total; we evaluate on the 1,533-sample test split). A comprehensive spatial reasoning benchmark with 50 fine-grained subcategories organized into four major categories: dynamic reasoning (e.g., motion analysis, collision prediction), complex spatial logic (e.g., multi-step reasoning), spatial interaction (e.g., navigation, traffic analysis), and perspective-taking (e.g., viewpoint changes). The test split is disjoint from the training data used in our experiments. *Example:* "How long will it take for the moving car closest to the camera to reach it if it is going at 10 m/s?"

D.1.3 Benchmark Metrics

Table 4 specifies the evaluation scope and metric for every main-table benchmark. Scores are reported as percentages, and higher is better. Unless stated otherwise, the benchmark parser extracts a canonical answer and reports exact answer or option accuracy; an extraction failure receives zero credit.

Table 4: Evaluation scope and metrics for the 13 benchmarks.

Benchmark	Evaluated scope	Reported metric
VSI-Bench	2,490 MC + 2,640 regression	Equal average of MC accuracy and regression mean relative accuracy (MRA).
VideoMMMU	882 MC + 18 numerical	Mean per-question score: exact option accuracy for MC and normalized numerical exact match for numerical questions.
MMVU (MC)	625 validation MC questions	Exact option accuracy; the 375 validation open-ended questions are excluded.
MVBench	4,000 MC questions, 20 tasks	Exact option accuracy over all questions.
TempCompass	7,540 converted MC questions	Exact option accuracy over all questions.
VideoMME	2,700 MC questions	Exact option accuracy over all questions.
CV-Bench	1,438 2D + 1,200 3D questions	Exact option accuracy over the combined 2D and 3D subsets.
SPAR-Bench	7,211 released test records	Mean per-question task score: MRA for numerical depth/distance tasks and canonical answer accuracy otherwise.
ViewSpatial-Bench	5,712 questions	Exact canonical answer accuracy.
MMSI-Bench	1,000 MC questions	Exact option accuracy.
MindCube	21,154 questions	Exact canonical answer accuracy.
OST-Bench	3,194 reformulated single-turn questions	Mean task score using option judgement, relative-error estimation, or exact enumeration according to question type.
OmniSpatial	1,533 test questions	Exact canonical answer accuracy.

For VideoMMMU numerical questions, predictions are normalized as numbers and compared after rounding to two decimal places. VSI-Bench regression, SPAR numerical questions, and OST

estimation use ten relative-error thresholds from 0.50 to 0.95, yielding fractional per-item credit in increments of 0.1; their precise boundary and zero-target conventions are given in Appendix E.2. The final SPAR and OST scores average their task-specific per-question scores rather than binarizing fractional credit. *Average Score* in the ablation table is the unweighted mean of the 13 full-benchmark scores.

The SPAR-Bench paper and repository describe the finalized benchmark as containing 7,207 QA pairs, whereas the currently distributed Hugging Face test split contains 7,211 records. Our evaluation follows the distributed split; this four-record difference reflects the released dataset version rather than additional filtering by our pipeline.

OST-Bench is originally organized as 10,165 QA turns in multi-turn scene dialogues. We reformulate 3,194 questions as independent single-turn instances for evaluation. All in-house OST results in the main table use this scope and therefore do not represent performance under the original multi-turn protocol.

These benchmark scores are distinct from the paired duality diagnostics. Accuracy/MRA evaluates each answer against its task target, whereas consistency evaluates whether an original–dual pair obeys its declared answer map. Appendix E.2 defines consistency, joint correctness, wrong-consistency, and the treatment of score-thresholded numerical answers.

D.2 Baseline Result Sources and Evaluation Settings

Our comparisons combine literature-reported scores and our own evaluations because reproducing every baseline is computationally expensive. We distinguish the source of each model–benchmark result: a literature-reported score is taken from a cited paper, whereas a reproduced score is obtained by running the corresponding model in our evaluation pipeline. Reproduced evaluation does not necessarily imply retraining the baseline. Scores from different sources may use different inference settings and are not treated as controlled comparisons.

Result provenance. Table 5 summarizes result sources. Attribution is made per model–benchmark entry: reproducing one benchmark does not change the provenance of the remaining reported scores in that row. The following paragraphs give source-specific settings and coverage.

Frontier video baseline provenance. All newly added GPT-5, GPT-5.4, and Gemini-3-Pro-Preview scores in Table 1 are literature-reported, not our reproductions. For GPT-5, VSI-Bench 55.0 comes from SenseNova-SI (arXiv:2511.13719v2, Table 1; GPT-5-2025-08-07) Cai et al. (2026); VideoMMMU 84.6 and MMVU (MC) 82.6 come from Video-KTR Table 1 Wang et al. (2026); MVBench 74.1 and VideoMME 83.3 come from CoPE-VideoLM Tables 3 and 2, respectively, with the latter using no subtitles. For GPT-5.4, VideoMMMU 87.1 and MVBench 78.3 come from VideoKR Table 3 (64 frames).

For Gemini-3-Pro-Preview, VSI-Bench 52.5 is explicitly attributed to the Preview variant in SenseNova-SI Table 1, and VideoMMMU 87.6 is also reported in VideoKR Table 3 for the November 2025 Gemini 3 Pro model. The Preview designation follows the explicitly named SenseNova-SI entry; VideoKR shortens the model name to Gemini 3 Pro and does not specify an exact API snapshot. Other scores labelled only Gemini 3 Pro are withheld pending version verification. MMVU scores with an unresolved match to our MC split, TempCompass scores with an unresolved match to our converted evaluation set, and GPT-5.4 VideoMME with unverified subtitle settings remain unfilled. Each populated cell retains its source's evaluation settings; an unspecified API snapshot or inference parameter is not inferred from another source.

Spatial result provenance. VST-7B-RL uses Qwen2.5-VL-7B without an additional specialist 3D encoder. Its CV-Bench (86.5) and MMSI (34.8) scores are reported values from Table 2 of Yang et al. Yang et al. (2025c) (arXiv:2511.05491v1), where neither is designated as a restricted subset; the other five populated cells are our evaluations. VST uses 4.1M perception samples, 135K reasoning samples, and additional general multimodal data, so it is not a training-budget-matched comparison.

For GPT-4o, InternVL-2.5-8B, LLaVA-OV-7B, and SpatialLadder-3B, the first five benchmark scores in Table 2 retain their literature attribution to SpatialLadder Li et al. (2026b); populated OST and OmniSpatial cells are our evaluations. For SenseNova-SI-8B, ViewSpatial and MMSI are reported

Table 5: Result provenance for the main comparison tables.

Models	Source	Scope / reference
Qwen2.5-VL-7B-Instruct and +SAGE	Our evaluations	Main video and spatial results; settings in Appendix D.3.
Qwen3-VL-4B/8B-Instruct and +SAGE	Our evaluations	Both backbone and +SAGE scores cover all six video and seven spatial benchmarks.
Video-R1, Video-RTS, VideoRFT, ViSS-R1	Reported	Video results attributed to the respective papers Feng et al. (2025); Wang et al. (2025c;b); Fang et al. (2025).
Proprietary VLMs and additional video baselines	Reported	Source-specific settings and coverage are documented below and in Table 14.
InternVL-2.5-8B, LLaVA-OV-7B, and SpatialLadder-3B	Reported and our evaluations	The first five spatial benchmarks use values reported by SpatialLadder Li et al. (2026b); OST and OmniSpatial are our evaluations.
Video-RAISE-7B	Reported	Collected from Table 1 of Xi et al. Xi et al. (2026), using the 16-frame row.
Video-KTR-7B	Reported and our evaluations	VideoMMMU, MMVU, TempCompass, and VideoMME are from Table 1 of Wang et al. Wang et al. (2026), using the 16-frame row; VSI-Bench and MVBench are our evaluations.
VideoSSR, VideoAuto-R1, and One-Thinker	Reported and our evaluations	Literature-reported cells and inference settings appear in Table 6; the remaining populated cells are our evaluations.
VST-7B-RL	Reported and our evaluations	CV-Bench and MMSI are from Table 2 of Yang et al. Yang et al. (2025c); the remaining populated cells are our evaluations.
SenseNova-SI-8B	Reported and our evaluations	InternVL3-8B variant; ViewSpatial and MMSI are from Table 1 of Cai et al. Cai et al. (2026), and the remaining populated cells are our evaluations. MindCube-Tiny is excluded.

values, while the other populated cells are our evaluations. SenseNova-SI-8B denotes the InternVL3-8B variant, not a Qwen3 checkpoint. All Qwen2.5-VL-7B-Instruct and Qwen3-VL-4B/8B-Instruct rows, including their +SAGE variants, are from our evaluations; each improvement is computed only against its matched same-backbone baseline.

Source documentation and benchmark coverage. Table 6 documents nine literature-reported scores and their source-specific inference settings. A single frame configuration is retained per method; these are contextual comparisons, not compute-matched runs. The remaining populated cells for VideoSSR, VideoAuto-R1, and OneThinker, together with all Qwen3-VL-4B/8B-Instruct and +SAGE rows, are from our evaluations.

OneThinker's released VideoMMMU/MMVU/VideoMME manifests contain 900/625/2,700 examples and match our problem-ID sets. VideoAuto-R1's reported configuration covers all VideoMMMU tasks, the 625-question MMVU MC validation split, all 20 MVBench tasks, and VideoMME, without a sample limit. VideoSSR reports standard MVBench and VideoMME results; its evaluator excludes missing or failed samples. Its other populated cells use our evaluation records rather than the previously inspected MC-only VideoMMMU result.

MMVU split and reporting scope. MMVU denotes all 625 multiple-choice questions in the public validation split, excluding the 375 open-ended questions. Gemini-2.0-Flash's previously collected 66.5 is an overall test score covering both question formats Zhao et al. (2025c), not a matching MC result, so its entry remains "-". VideoSSR's 71.2 is our reproduced validation-MC result. Missing

Table 6: Sources and inference settings of reported Qwen3-based baseline results.

Method	Reported scores / source	Source inference settings
VideoSSR-8B	MVBench 68.9; VideoMME 67.6. He et al., *VideoSSR: Video Self-Supervised Reinforcement Learning*, Table 2 (v1).	64-frame row; 2 FPS; 512×512 pixel budget; greedy direct answering without CoT; VideoMME without subtitles.
VideoAuto-R1-8B	VideoMMMU 65.0; MMVU 71.0; MVBench 71.8; VideoMME 67.9. Liu et al., *VideoAuto-R1: Video Auto Reasoning via Thinking Once, Answering Twice*, CVPR supplemental Table 7.	64-frame row; confidence-based early exit; greedy decoding; VideoMME without subtitles. The selected row differs from the released launcher's larger default frame limits.
OneThinker-8B	VideoMMMU 66.2; MMVU 70.5; VideoMME 66.5. Feng et al., *OneThinker: All-in-one Reasoning Model for Image and Video*, Table 2 (v1).	Up to 128 frames; 2 FPS; per-frame pixel cap 262,144; greedy decoding; released evaluator uses no VideoMME subtitles.

results are marked "-" rather than substituted from a different split. These exclusions do not imply reporting errors in the source papers.

Video-KTR (65.7; Table 1, 16 frames) Wang et al. (2026) and OneThinker (70.5; Table 2) Feng et al. (2026) explicitly report MMVU(mc), so their values are retained as published. Their reported decimals do not match a single-run accuracy rounded to one decimal place over 625 questions. The effective denominator, aggregation, and reporting convention remain unresolved; we preserve the published values without attributing the discrepancy to an error in either paper.

Literature-reported scores retain their documented model variants, metrics, splits, and inference settings. In-house results use checkpoint-specific evaluation records. MindCube-Tiny and other restricted subsets are not substitutes for the full evaluation splits used here.

D.3 Training Configuration

Shared-model hyperparameters. Table 7 summarizes the shared-model training configuration. Proposer-format SFT uses a learning rate of 10^{-5} for one epoch, followed by the RL configuration specified in Section 3.1.

The probe buffer contains up to 200 source items from training rollouts; only applicable pairs count toward $N_{\min}$. Profiles are initialized from admission probes. For an operation with raw probe estimate $\widehat{z}_i$, each stored score $z_i \in \{U_i, J_i\}$ is updated by

$$z_i \leftarrow \begin{cases} 0.9z_i + 0.1\widehat{z}_i, & N_i \geq N_{\min}, \\ z_i, & N_i < N_{\min}. \end{cases} \tag{20}$$

Thus, empty or undersupported probes do not reset an operation's scores. The proposer baseline is initialized as $b_0 = 0$. Retirement requires $U \leq 0.05$ and $J \geq 0.90$; reactivation occurs when $U > 0.10$ or $J < 0.80$.

Model initialization and data. Qwen2.5-VL-7B-Instruct and Qwen3-VL-4B/8B-Instruct receive proposer-format SFT on 1,000 examples selected from 2,000 duality-operation code examples generated by DeepSeek-V4-Pro, followed directly by GRPO, without a Video-R1 task-SFT warm-up. The RL dataset contains 10K samples, comprising 5K from Video-R1 and 5K from SpatialLadder. Training uses 16 NVIDIA A100 GPUs, DeepSpeed ZeRO-3, 8 original rollouts and 4 dual rollouts when applicable, for one epoch.

Training-data scale. For context, Video-R1 constructs a 260K-sample RL corpus, while SpatialLadder trains progressively on its 26,610-sample spatial corpus Feng et al. (2025); Li et al. (2026b); SAGE uses 10K samples for RL.

Optimization details. We use the framework-default AdamW optimizer with a cosine learning-rate schedule. The learning rate is 10^{-6}, with weight decay 0.01, maximum gradient norm 5, per-device

Table 7: Hyperparameters for shared-model SAGE training.

Hyperparameter	Value
Proposal interval E_p (solver steps)	100
Proposer update weight η	0.01
KL coefficient β	0.04
Mastered-operation spot-check probability p_f	0.20
Proposal-baseline EMA momentum ρ	0.9
Operation-profile EMA momentum	0.9
Maximum pool capacity M	50
Probe buffer capacity (source items)	200
Maximum proposal output tokens	1,024
Maximum probe-answer output tokens	128
Advantage normalization ϵ	10^{-4}
Minimum applicable pairs $N_{\min}$	32
Admission threshold τ_{add}	0.10
One-sided-failure thresholds $(\tau_{\text{lo}}, \tau_{\text{hi}})$	$(0.05, 0.10)$
Joint-correctness thresholds $(\kappa_{\text{lo}}, \kappa_{\text{hi}})$	$(0.80, 0.90)$
GRPO ratio-clipping half-width	0.20
Gradient accumulation steps	1
Maximum gradient norm	5.0
Maximum prompt / completion length	16,384 / 768
Rollout temperature / top-p	1.0 / 0.95
Proposer-format SFT learning rate	10^{-5}
Proposer-format SFT epochs	1
Proposer-format SFT examples	1,000

batch size 1, and one gradient accumulation step. Maximum prompt and completion lengths are 16,384 and 768 tokens, respectively. Rollouts use temperature 1.0 and top-$p = 0.95$. Training uses bf16, gradient checkpointing, and Flash Attention 2, with a maximum visual input budget of 401,408 pixels. We set the KL coefficient to $\beta = 0.04$ and the mastered-operation spot-check probability to $p_f = 0.2$.

Evaluation protocol. Literature-reported scores retain their source-specific inference settings, documented in Appendix D.2. The recorded in-house configuration specifies greedy decoding (temperature ≈ 0, top-p $= 0.001$), at most 1,024 output tokens, and 16 frames per video.

E Additional Experimental Results

This appendix provides controlled ablation protocols, paired-metric definitions, and additional analyses of training and operation-pool behavior.

E.1 Controlled Ablation Protocols

All controlled variants use the same backbone, proposer-format initialization, training data, and evaluation protocol unless the named component is changed. Average Score is the unweighted mean of the 13 full-benchmark scores. Paired diagnostics use a common evaluation set disjoint from the training probe buffer.

The ablation backbone is Qwen2.5-VL-7B-Instruct. Base and full-SAGE averages are calculated from the corresponding rows of Tables 1 and 2, using their displayed benchmark scores and rounding only after aggregation. The top SAGE row is the complete configuration shared by the relevant panel-wise comparisons.

Duality supervision. Table 3 compares paired training on a fixed stream of valid original–dual examples with matched transformations, labels, and rollout budgets. Dual augmentation uses task and format rewards without a consistency bonus, updating first on the original input and then on the dual input. The ungated consistency variant removes the correctness gate, whereas SAGE retains it and

combines both sides' gradients before one solver update. Thus comparison with dual augmentation changes both reward design and update scheduling; Appendix F.5 isolates the latter with the loss components held fixed. Under the SAGE reward, one-sided-correct pairs retain task supervision but receive no consistency bonus. Original-only GRPO is a reference, not an equal-rollout-cost control.

Operation scheduling and discovery. In Panel C of Table 3, Fixed-all applies every applicable operation, while Random-K randomly selects at most K. Adaptive-K balances operation usage, prioritizing less-used operations to provide comparable training exposure. It does not use probe results or consistency scores to set priorities. Random-K and Adaptive-K share the active-capacity limit and solver rollout budget. Panel D removes one mechanism from SAGE at a time: removing discovery retains the initial operations and probe-based management; removing retirement disables Active-to-Mastered transitions; removing reactivation disables re-entry of mastered operations.

Operation families. Table 3 restricts admissible operations to visual or linguistic families under the same training recipe and rollout budget; their combination is represented by the top SAGE reference row.

Shared-model updates. Table 3 isolates parameter sharing and solver-priority projection. The unprojected variant retains the same proposer weight and objectives; the separate proposer–solver uses two trainable role models initialized from the same backbone. The table reports performance and resource use under matched rollout budgets.

E.2 Paired-Metric Definitions

Task scores versus binary correctness. For each of the N evaluated pairs, let $s_i, s'_i \in [0,1]$ be the original and dual task scores returned by the benchmark scorer, and let c_i indicate agreement under the declared answer map. The paired scores are

$$A_{\text{orig}} = \frac{1}{N}\sum_i s_i, \qquad A_{\text{dual}} = \frac{1}{N}\sum_i s'_i, \qquad C = \frac{1}{N}\sum_i c_i. \tag{21}$$

Original-side accuracy is A_{orig}, not the mean over both sides. For score-thresholded diagnostic profiles, $b_i = \mathbf{1}[s_i > 0.5]$ and $b'_i = \mathbf{1}[s'_i > 0.5]$; a score of exactly 0.5 counts as wrong. Joint correctness $J = N^{\text{CC}}/N$ is distinct from the average task score.

Correctness-conditioned consistency. Joint correctness and wrong-consistency are

$$J = \frac{1}{N}\sum_i b_i b'_i, \qquad C_{\text{wrong}} = \frac{1}{N}\sum_i (1-b_i)(1-b'_i)c_i. \tag{22}$$

Under exact canonical correctness and a valid bijective answer map, both-correct pairs are consistent and one-sided-correct pairs are inconsistent, giving $C = J + C_{\text{wrong}}$. Score-thresholded numerical diagnostics need not satisfy this identity: two numerical answers may both pass the task threshold without obeying the declared answer relation. Correct-consistency therefore measures $N^{-1}\sum_i b_i b'_i c_i$, not joint correctness. The proofs in Appendix F use exact canonical correctness, not thresholded numerical scores.

Numerical scoring and consistency. The repository's VSI-Bench regression, SPAR numerical, and OST estimation scorers use ten relative-error thresholds spanning 0.50 to 0.95, giving per-item credit in increments of 0.1. Their boundary conventions differ: video regression and OST estimation use strict inequalities, whereas SPAR uses a non-strict inequality; zero-target handling also differs. The paired evaluation implementation separately normalizes numerical answers to two decimal places before testing their mapped equality. This answer-equality test is not the fractional-credit task score. Recomputations therefore retain the original per-item scores and normalized mapped answers rather than replacing benchmark-specific scorers with a shared tolerance.

Aggregation and numerical precision. Paired diagnostics pool valid applicable pairs rather than averaging benchmark-level percentages. All aggregation precedes rounding to one decimal place. Full-benchmark scores retain their benchmark-specific aggregation, and Average Score averages the 13 completed benchmark scores equally. Missing benchmarks are not treated as zero or silently dropped.

Table 8: Checkpoint-level evolution of correctness and duality consistency.

Checkpoint	Average Score	Consistency	Joint Correctness	Wrong-consistent
Base	43.8	61.4	24.1	37.3
Early	49.2	70.7	37.8	32.9
Middle	51.5	74.1	42.9	31.2
Final	52.9	75.9	45.7	30.2

Table 9: Generalization to unseen duality operations.

Method	Seen Cons.	Unseen Cons.	Seen J	Unseen J	ΔJ
Base	61.67	56.64	24.00	22.31	-1.69
GRPO	68.50	62.89	35.95	32.51	-3.44
SAGE	76.25	73.02	45.62	43.11	-2.51

Table 10: Quality of generated duality operations before pool admission.

Model / Stage	Valid code (%)	Executable (%)	Applicable (%)	Correct mapping (%)	Accepted (%)
Initial proposals	91.67	81.67	65.00	53.33	16.67
After training	96.67	93.33	93.33	88.33	46.67

E.3 Further Analysis Details

Checkpoint-level dynamics. Table 8 uses the same metrics as Table 3, with Base and Final aligned to its reference rows. Average Score averages the 13 benchmarks; the paired diagnostics distinguish joint correctness from wrong-consistent predictions.

Generalization to unseen operations. The held-out transformations are never activated during training and are evaluated without further model updates. Reporting both consistency and joint correctness distinguishes transferable reasoning from merely agreeing under unfamiliar transformations.

Proposal quality. Tables 10 and 11 assess parsing and execution, applicability, answer-map correctness, and pool admission. Comparing generated proposals with the initial pool further distinguishes valid operations from novel and useful ones.

Reward weight and active-set size. Tables 12 and 13 vary λ and K to examine the balance between task performance, consistency, and operation coverage while keeping the remaining settings fixed.

Limitations and discussion. Consistency does not imply correctness: two wrong predictions can still obey the answer mapping. Correctness gating reduces this training incentive but does not eliminate wrong-consistent predictions at evaluation. Executable checks cannot guarantee semantic validity for all ambiguous operations, particularly color, orientation, viewpoint, and negation transformations. SAGE also incurs additional rollout computation and remains dependent on backbone capability. Results are primarily based on single runs; multiple-seed experiments and broader backbone comparisons remain necessary.

E.4 Additional Video Baselines

Table 14 extends the comparison to earlier and general-purpose video VLMs whose reported coverage is incomplete across the six benchmarks. Their strongest results are concentrated on individual benchmarks, so we retain these sparsely populated rows as contextual references rather than use them for aggregate comparison.

Table 11: Novelty and quality of generated duality operations.

Proposal source	Total	Valid (%)	Novel (%)	Useful (%)
Initial pool	6	100.00	–	100.00
Generated proposals	60	88.33	70.00	56.67

Table 12: Sensitivity to the duality reward weight λ.

λ	Average Score	Consistency	Joint Correctness	Wrong-consistent
0.0	51.6	74.9	43.9	31.0
0.1	52.5	75.6	45.0	30.6
0.3	52.9	75.9	45.7	30.2
0.5	52.8	76.3	45.6	30.7
0.7	52.2	76.8	45.0	31.8
1.0	51.5	77.0	44.0	33.0

Table 13: Sensitivity to the number of active duality operations.

K	Average Score	Consistency	Joint Correctness	Wrong-consistent
1	51.6	73.6	43.5	30.1
2	52.6	75.3	45.2	30.1
3	52.9	75.9	45.7	30.2
4	53.0	76.0	45.9	30.1
6	52.7	75.7	45.4	30.3

Table 14: Additional third-party video results omitted from the compact main table. Values are quoted from the cited papers.

Model	VSI-Bench	VideoMMMU	MMVU	MVBench	TempCmps	VideoMME
LLaMA-VID Li et al. (2025c)	-	-	-	41.9	45.6	-
VideoLLaMA2 Cheng et al. (2024b)	-	-	44.8	54.6	-	47.9
LongVA-7B Zhang et al. (2025a)	29.2	23.9	-	-	56.9	52.6
VILA-1.5-40B Lin et al. (2024)	31.2	34.0	-	-	-	60.1
Video-UTR-7B Yu et al. (2025)	-	-	-	58.8	59.7	52.6
LLaVA-OneVision-7B Li et al. (2025a)	32.4	33.8	49.2	56.7	-	58.2
Kangaroo-8B Liu et al. (2026a)	-	-	-	61.1	62.5	56.0

E.5 Training Dynamics

Figure 5 shows an early decline in consistency followed by recovery, while accuracy and total rewards remain stable. The bounded KL divergence is consistent with stable policy adaptation over the observed run.

F Properties of Answer-Level Equivariance

We state only the properties needed by SAGE. We do not claim a VC-dimension or generalization bound for an overparameterized VLM.

F.1 Setup

We use the canonical decoder $\hat{a}_\theta$ from Section 2.1 and deterministic ground truth $a^*(x)$. This analysis uses exact canonical 0–1 correctness; task-specific tolerant reward verification is outside the theorem. For a valid $\mathcal{T} = (T_\mathcal{X}, T_\mathcal{A}, \mathcal{S})$ with $\Pr_{x\sim\mathcal{D}}[x \in \mathcal{S}] > 0$, all probabilities below are over $x \sim \mathcal{D}_\mathcal{S}$, and

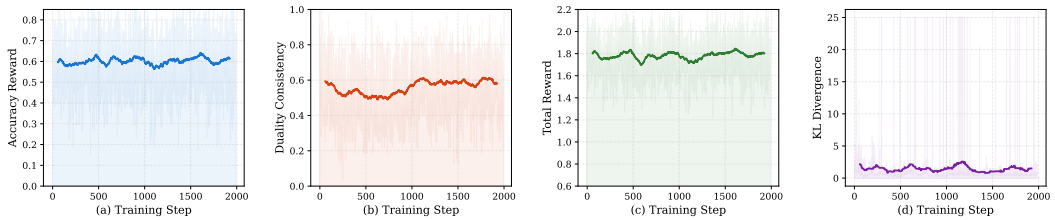

Figure 5: Training dynamics of SAGE. Accuracy reward remains stable, duality consistency recovers after an initial change, total reward remains stable, and KL divergence stays bounded.

extraction failure $\perp$ counts as an incorrect answer. Define

$$\begin{aligned} R(\theta) &= \Pr[\hat{a}_\theta(x) \neq a^*(x)]\,, \\ R^{\mathcal{T}}(\theta) &= \Pr[\hat{a}_\theta(T_{\mathcal{X}}(x)) \neq T_{\mathcal{A}}(a^*(x))]\,, \\ R_{\text{pair}}(\theta) &= \tfrac{1}{2}\left(R(\theta) + R^{\mathcal{T}}(\theta)\right). \end{aligned} \tag{23}$$

F.2 Effect of Group-Relative Normalization

The correctness gate makes the consistency term a cross-side reward reweighting rather than an additional source of labels. Let $x_i \in \{0,1\}$ and $y_j \in \{0,1\}$ denote exact correctness of the original and dual completions. By bijectivity and Eq. (1), $g_{ij}c_{ij} = x_i y_j$, and hence

$$r_{\text{cons}}(o_i) = \lambda x_i \bar{y}, \qquad \bar{y} = \frac{2}{G}\sum_{j=1}^{G/2} y_j, \tag{24}$$

with the symmetric identity $r'_{\text{cons}}(o'_j) = \lambda y_j \bar{x}$ on the dual side, where $\bar{x} = G^{-1}\sum_{i=1}^{G} x_i$.

For a reward vector z and $\epsilon \geq 0$ with $\sigma_z + \epsilon > 0$, define its group-relative normalization by

$$\mathsf{Z}_\epsilon(z)_i = \frac{z_i - \bar{z}}{\sigma_z + \epsilon}. \tag{25}$$

Proposition 1 (Affine-invariant boundary case)**.** *Let the original-side reward before the consistency term be $q_i = ax_i + c$, where $a > 0$ and c is constant within the rollout group. Then adding Eq. (24) produces a positive affine transformation of q. If $\sigma_q > 0$, then $\mathsf{Z}_0(q + r_{\text{cons}}) = \mathsf{Z}_0(q)$; for $\epsilon > 0$, the two normalized vectors are positive scalar multiples of one another. If $\sigma_q = 0$ and $\epsilon > 0$, both normalized vectors are zero. Outside this affine condition, group normalization provides no such invariance guarantee.*

Proof. Set $k = \lambda\bar{y} \geq 0$. Since $q_i = ax_i + c$,

$$q_i + kx_i = \left(1 + \frac{k}{a}\right) q_i - \frac{kc}{a} = Aq_i + B, \tag{26}$$

where $A = 1 + k/a > 0$ and $B = -kc/a$. Thus $\overline{Aq+B} = A\bar{q} + B$ and $\sigma_{Aq+B} = A\sigma_q$. Substitution into Eq. (25) proves exact invariance for $\epsilon = 0$ and positive collinearity for $\epsilon > 0$. When $\sigma_q = 0$, x_i is constant because $a > 0$, so the augmented rewards are also constant. If the base reward is not affine in binary correctness, adding kx is not generally a positive affine transformation of the reward vector, and the preceding invariance argument does not apply. □

Binary accuracy with a group-constant format reward is the principal boundary case covered by Proposition 1. Graded task rewards or completion-dependent format rewards can violate its affine premise, allowing cross-side reweighting to change relative advantages. The proposition therefore identifies when the bonus is redundant under normalization; it does not attribute gains from paired task supervision, joint updating, or adaptive operation selection to the bonus alone.

F.3 Pair-Risk Lower Bound

Theorem 1 (Pair-risk lower bound)**.** *For any valid duality operation,*

(i) $R_{\text{pair}}(\theta) \geq (1 - \mathcal{C}_{\mathcal{T}}(\theta))/2$*;*

(ii) if $R(\theta) = 0$*, then* $R^{\mathcal{T}}(\theta) = 1 - \mathcal{C}_{\mathcal{T}}(\theta)$*.*

Proof. Fix $x \in \mathcal{S}$ and write $a = \hat{a}_\theta(x)$ and $a' = \hat{a}_\theta(T_{\mathcal{X}}(x))$. Let

$$I_{\text{cons}}(x) = \mathbf{1}[a \in \mathcal{A} \ \wedge \ a' \in \mathcal{A} \ \wedge \ T_{\mathcal{A}}(a) = a'] \,. \tag{27}$$

If both predictions are correct, then $a = a^*(x)$ and, by Eq. (1), $a' = T_{\mathcal{A}}(a^*(x)) = T_{\mathcal{A}}(a)$, so $I_{\text{cons}}(x) = 1$. Taking the contrapositive, every inconsistent sample has at least one incorrect side, including any sample with an extraction failure. Hence

$$\mathbf{1}[a \neq a^*(x)] + \mathbf{1}[a' \neq T_{\mathcal{A}}(a^*(x))] \geq 1 - I_{\text{cons}}(x). \tag{28}$$

Taking expectations and dividing by two proves (i). If $R(\theta) = 0$, then $a = a^*(x)$ almost surely. On those samples, $I_{\text{cons}}(x) = 1$ if and only if $a' = T_{\mathcal{A}}(a^*(x))$; therefore $R^{\mathcal{T}}(\theta) = 1 - \mathcal{C}_{\mathcal{T}}(\theta)$, proving (ii). Perfect consistency alone does not imply zero pair risk: a consistent model may be wrong on both sides. This is precisely why Eq. (14) is correctness-gated. □

F.4 Noise Robustness of Hysteretic Pool Updates

Proposition 2. *Let* $U_t^\star, J_t^\star$ *denote the noiseless operation scores and let the transition statistics be* $\widehat{U}_t = U_t^\star + \epsilon_t^U$ *and* $\widehat{J}_t = J_t^\star + \epsilon_t^J$*, where* $|\epsilon_t^U| \leq \delta_U$ *and* $|\epsilon_t^J| \leq \delta_J$*. Suppose Eqs. (9)–(10) are applied to* $\widehat{U}_t, \widehat{J}_t$ *and*

$$\tau_{\text{hi}} - \tau_{\text{lo}} > 2\delta_U, \qquad \kappa_{\text{hi}} - \kappa_{\text{lo}} > 2\delta_J. \tag{29}$$

After an Active-to-Mastered transition at refresh t_1*, a later Mastered-to-Candidate transition at* t_2 *requires at least one of*

$$U_{t_2}^\star - U_{t_1}^\star > \tau_{\text{hi}} - \tau_{\text{lo}} - 2\delta_U, \tag{30}$$

$$J_{t_1}^\star - J_{t_2}^\star > \kappa_{\text{hi}} - \kappa_{\text{lo}} - 2\delta_J. \tag{31}$$

Thus bounded measurement noise alone cannot cause a reversal when the underlying scores are unchanged.

Proof. Mastery at t_1 implies $U_{t_1}^\star \leq \tau_{\text{lo}} + \delta_U$ and $J_{t_1}^\star \geq \kappa_{\text{hi}} - \delta_J$. Reactivation occurs through one of the two conditions in Eq. (10). If $\widehat{U}_{t_2} > \tau_{\text{hi}}$, then $U_{t_2}^\star > \tau_{\text{hi}} - \delta_U$; subtracting the bound at t_1 gives Eq. (30). If $\widehat{J}_{t_2} < \kappa_{\text{lo}}$, then $J_{t_2}^\star < \kappa_{\text{lo}} + \delta_J$; subtracting from the bound at t_1 gives Eq. (31). Both lower bounds are positive by assumption. □

F.5 Joint Versus Sequential Paired Updates

Comparison and potential advantage. SAGE evaluates the original and dual losses at the same parameter state and combines their gradients before a solver update. The sequential dual-augmentation comparator instead updates on the original input and then on the dual input at the intermediate parameters. Joint updating avoids this within-pair parameter drift and uses one rather than two optimizer steps. The following analysis isolates update scheduling: both methods optimize the same two loss components. It does not change SAGE's reward or replace it with a single all-or-nothing pair reward.

Let $\ell_o(\theta)$ and $\ell_d(\theta)$ be fixed differentiable original and dual loss components, and let $L(\theta) = \frac{1}{2}(\ell_o(\theta) + \ell_d(\theta))$. Match first-order gradient weights by comparing one joint step of size α with two sequential steps of size $\alpha/2$:

$$\theta^+_{\text{joint}} = \theta - \frac{\alpha}{2}\big(\nabla \ell_o(\theta) + \nabla \ell_d(\theta)\big), \tag{32}$$

$$\theta_{1/2} = \theta - \frac{\alpha}{2}\nabla \ell_o(\theta), \qquad \theta^+_{\text{seq}} = \theta_{1/2} - \frac{\alpha}{2}\nabla \ell_d(\theta_{1/2}). \tag{33}$$

For a locally Lipschitz Hessian of ℓ_d, Taylor expansion as $\alpha \to 0$ yields

$$\theta^+_{\text{seq}} - \theta^+_{\text{joint}} = \frac{\alpha^2}{4}\nabla^2\ell_d(\theta)\nabla\ell_o(\theta) + O(\alpha^3). \tag{34}$$

Thus, sequential updating introduces an order-dependent second-order term absent from joint updating. Its sign is not universally unfavorable; the proposition below gives explicit conditions for a strict joint-update advantage.

Proposition 3 (Eliminating sequential residual error under conflicting objectives)**.** *Let $\theta, a \in \mathbb{R}^d$, $a \neq 0$, and consider the equal-curvature quadratic losses*

$$\ell_o(\theta) = \tfrac{1}{2}\|\theta - a\|_2^2, \qquad \ell_d(\theta) = \tfrac{1}{2}\|\theta + a\|_2^2.$$

With exact gradients and constant $0 < \alpha < 2$, repeated joint updates in Eq. (32) converge from any finite initialization to the minimizer $\theta^* = 0$ of L. Original-then-dual sequential updates in Eq. (33), observed after each completed pair, converge to $-\alpha a/(4-\alpha)$. Their limiting excess average loss is

$$\lim_{t\to\infty}\left[L(\theta_t^{\text{seq}}) - L(\theta_t^{\text{joint}})\right] = \frac{\alpha^2\|a\|_2^2}{2(4-\alpha)^2} > 0. \tag{35}$$

Proof. The average loss is $L(\theta) = \frac{1}{2}\|\theta\|_2^2 + \frac{1}{2}\|a\|_2^2$. A joint step gives $\theta^{\text{joint}}_{t+1} = (1-\alpha)\theta^{\text{joint}}_t$, which converges to zero because $|1-\alpha| < 1$. A sequential cycle gives

$$\theta^{\text{seq}}_{t+1} = \left(1 - \frac{\alpha}{2}\right)^2\theta^{\text{seq}}_t - \frac{\alpha^2 a}{4}. \tag{36}$$

Its contraction factor lies in $(0,1)$, and its unique fixed point is $-\alpha a/(4-\alpha)$. Substituting this point and zero into L proves Eq. (35). Reversing the update order changes the sign of the fixed point but leaves the limiting excess loss unchanged. □

Interpretation and scope. The two objectives prefer different parameters, representing conflict between original and dual updates. Joint gradients balance these preferences at one parameter state, whereas cyclic sequential updates retain an order-dependent bias under constant steps. The proposition establishes a strict asymptotic loss advantage in this quadratic model, not superiority at every finite iteration or a guarantee of benchmark accuracy for a nonconvex VLM. Finite rollouts, GRPO clipping, momentum, AdamW preconditioning, and changing training samples are outside the proof. This is an advantage of update scheduling; it does not prove an independent benefit of the consistency reward. Dual augmentation that already accumulates both gradients before one optimizer step is not the sequential comparator.

Optimizer overhead and speed. Let C_{pair} be the common cost of generating and scoring both sides and computing their gradients, and let $U_{\text{opt}} > 0$ be the cost of one optimizer step, including its associated synchronization when applicable. Under this idealized decomposition,

$$T_{\text{joint}} = C_{\text{pair}} + U_{\text{opt}}, \qquad T_{\text{seq}} = C_{\text{pair}} + 2U_{\text{opt}}. \tag{37}$$

Joint updating saves one optimizer step per processed source pair. The modeled speedup is $T_{\text{seq}}/T_{\text{joint}} = 1 + U_{\text{opt}}/(C_{\text{pair}} + U_{\text{opt}})$, approaching one when generation and gradient computation dominate. Neither side's forward/backward work disappears, and communication overlap may differ in practice; a twofold end-to-end speedup is not implied.

Matched comparison protocol. The existing dual-augmentation and SAGE rows compare performance and resource use. A matched speed comparison uses identical hardware, paired inputs, rollout budgets, and timing boundaries. Full SAGE also incurs online proposal and probing overhead, so saving an optimizer step does not guarantee lower end-to-end training time. The theorem does not predetermine benchmark accuracy or measured runtime.

For the first-order-matched sequential control, each step uses half the joint learning rate; SAGE's existing learning-rate setting is unchanged. The schedule is indexed by processed source pairs, and per-pair regularization weights are matched rather than applying the full KL penalty twice. Both sides of a pair are generated before either update, retaining the same rollout budget. Rewards, advantages, and behavior-policy likelihoods are fixed within that pair, while the second sequential gradient is evaluated at the intermediate parameters. The proof holds losses fixed; comparisons that also change consistency rewards or online operation discovery measure their combined effect and cannot attribute the entire empirical gap to scheduling alone.

G COPYRIGHTS

We acknowledge that the image used in Figure 1 is credited to Kamran Aydinov on Magnific. The figure is used solely for academic illustration purposes.

H CASE STUDY: DUALITY CONSISTENCY IN PRACTICE

Visual operations. Figure 6 presents candidate diagnostic pairs from OmniSpatial. Rows 1–3 ask for quantities whose ground truth is preserved by the stated transformation, but the model's predictions change: a person's estimated speed changes from 0.43 m/s to 5.7 m/s after horizontal flip and to 24.7 m/s after 180° rotation; another speed changes from 3.5 m/s to 0.4 m/s after vertical flip; and a car-speed estimate changes from 2.33 m/s to 1.28 m/s after a flip plus color inversion. Row 4 is a boundary counterexample, not a valid consistency failure: color inversion changes the category-defining cue in "big black bear," so the applicability predicate must reject this pair.

Linguistic operations. Figure 7 shows failures under two linguistic operations. Under *option permutation*, the order of multiple-choice options is reversed while keeping the question and image unchanged. The model selects "B" in both cases, but "B" maps to "right" (correct) in the original and "left" (wrong) after permutation, indicating sensitivity to option position rather than consistent tracking of answer content. Under *question negation*, "Which hand does the woman use?" becomes "Which hand does NOT the woman use?" yet the model returns the identical answer, demonstrating a failure to track logical polarity in this example.

The valid pairs motivate the SAGE reward, while the rejected black-bear pair illustrates why every operation needs an applicability domain. The method applies consistency supervision only when the answer transformation is defined and the transformed sample remains semantically answerable.



Figure 6: Visual operation diagnostics. Rows 1–3 are answer-preserving pairs with inconsistent predictions (red). Row 4 is deliberately marked invalid: color inversion changes the category-defining "black" cue, so SAGE's applicability predicate excludes it from consistency supervision.

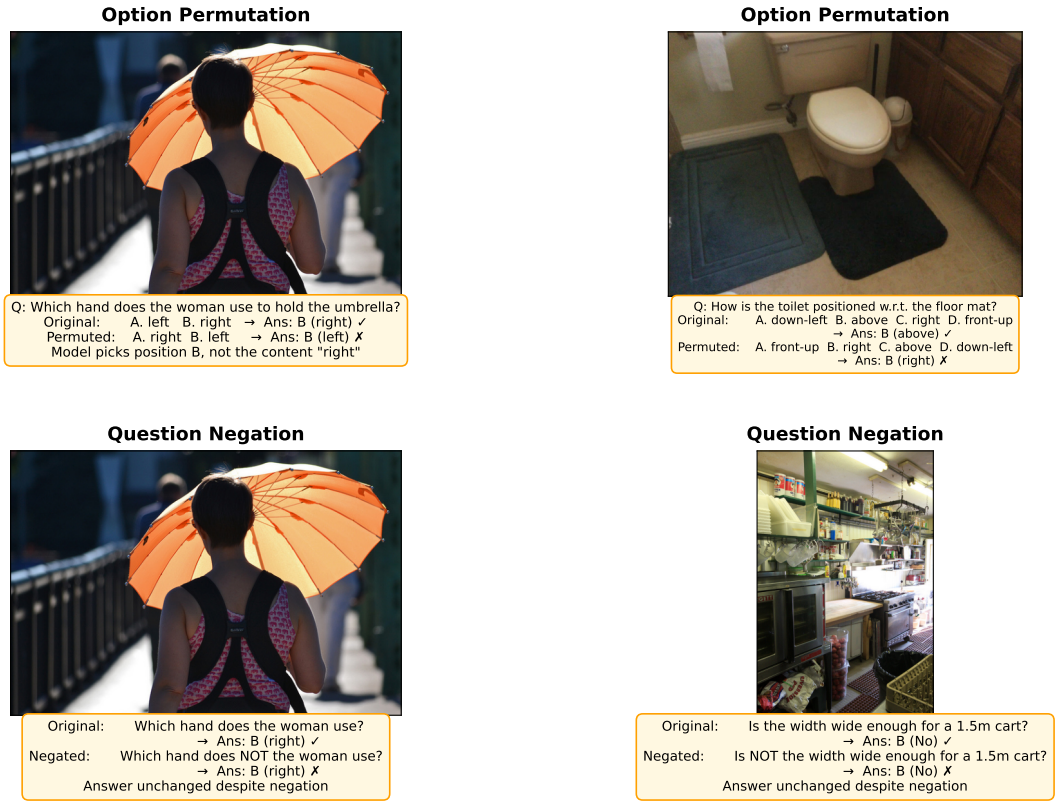

Figure 7: Linguistic duality operation failure cases. **Top**: option permutation reverses the order of choices; the model selects the same position (B) rather than tracking the content. **Bottom**: question negation inserts "NOT"; the model returns identical answers, ignoring the logical reversal.